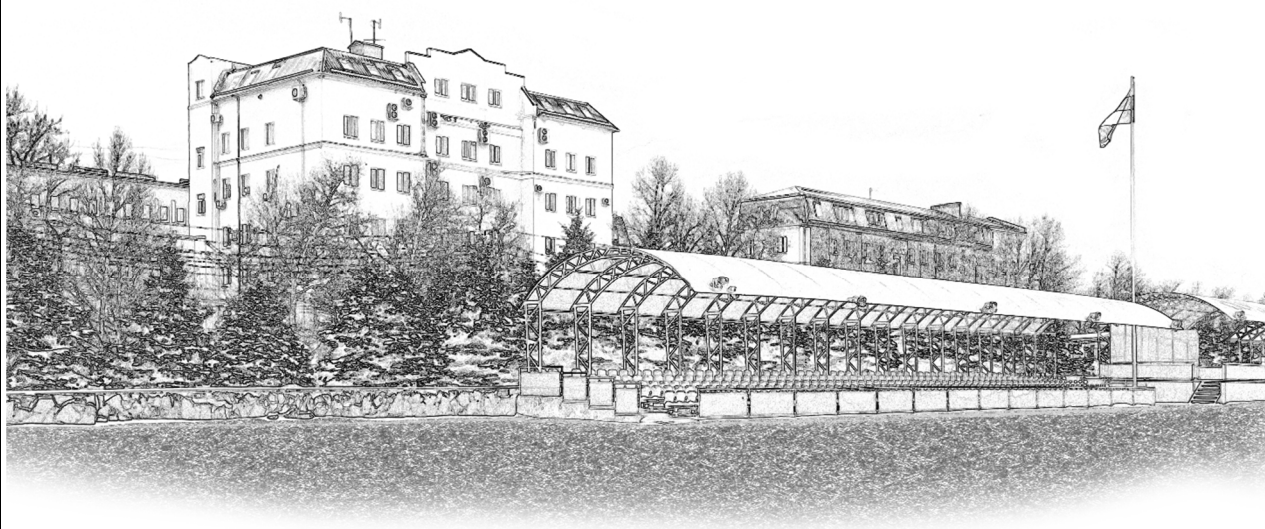




Краснодарский университет МВД России

ОБЕСПЕЧЕНИЕ НАДЕЖНОСТИ ТЕСТОВ ДЛЯ КОНТРОЛЯ ЗНАНИЙ



Краснодар
2026

Краснодарский университет МВД России

**ОБЕСПЕЧЕНИЕ НАДЕЖНОСТИ ТЕСТОВ
ДЛЯ КОНТРОЛЯ ЗНАНИЙ**

Краснодар
2026

УДК 378.146
ББК 74.48
О-136

Одобрено
редакционно-издательским советом
Краснодарского университета
МВД России

Рецензенты:

С. И. Агафонов, кандидат юридических наук, доцент (Московский университет МВД России имени В. Я. Кикотя);

Н. Ф. Гейжан, доктор педагогических наук, профессор (Санкт-Петербургский университет МВД России).

Авторы: *О. М. Булгаков, А. А. Хромых, Е. Н. Гривенная, А. О. Дедикова.*

Обеспечение надежности тестов для контроля знаний /
О-136 О. М. Булгаков и др. – Краснодар : Краснодарский университет
МВД России, 2026. – 180 с.

ISBN 978-5-9266-2247-5

В монографии систематизируется представление о дидактических возможностях фондов оценочных средств различных образовательных программ, которые в совокупности с доступностью программных оболочек с удобными интерфейсами представляют собой эффективный инструмент контроля учебных достижений обучающихся, мониторинга реализации образовательных программ, аудита качества образования. Исследуются особенности разработки компьютерных тестов и технологий их применения, особое внимание уделяется практическим рекомендациям по обеспечению достоверности оценивания учебных достижений обучающихся компьютерными тестами. Комплексно рассматриваются технологии тестирования учебных достижений в интересах эффективной подготовки сотрудников органов внутренних дел в ведомственных образовательных организациях.

Для профессорско-преподавательского состава, адъюнктов, курсантов, слушателей образовательных организаций МВД России и сотрудников органов внутренних дел Российской Федерации.

УДК 378.146
ББК 74.48

ISBN 978-5-9266-2247-5

© Краснодарский университет
МВД России, 2026

СПИСОК СОКРАЩЕНИЙ

ДЕСО – дидактическая единица содержания обучения.

ДОТ – дистанционные образовательные технологии.

ЕГЭ – единый государственный экзамен.

ИИ – искусственный интеллект.

КУДО – контроль учебных достижений обучающихся.

КТОЗУО – компьютерный тест для оценивания знаний и умений обучающихся.

СУО – система управления обучением.

ТЗ – тестовое задание.

ТЗАО – тестовое задание с альтернативными ответами.

ТЗВО – тестовое задание с выбором единственного верного ответа.

ТЗГО – тестовое задание с градацией ответов.

ТЗМВ – тестовое задание с множественным выбором верных ответов.

ТЗУД – тестовое задание с упорядочиванием дистракторов.

ТЗУП – тестовое задание на установление последовательности.

ТЗУС – тестовое задание на установление соответствия.

ТИИ – технологии искусственного интеллекта.

ТОЗУО – тест для оценивания знаний и умений обучающихся.

ТОУД – тест для оценивания учебных достижений.

ФОС – фонд оценочных средств.

ВВЕДЕНИЕ

Компьютерное тестирование учебных достижений обучающихся стало распространенным инструментом современной дидактики и педагогической квалиметрии, а сами тесты, как правило, являются важным компонентом ФОС различных образовательных программ. Благодаря своим дидактическим возможностям, доступности программных оболочек с удобными интерфейсами разработчика и испытуемого, упрощению организации сетевого и удаленного доступа КТОЗУО давно уже вышли из категории новых, нетрадиционных для отечественной педагогики дидактических средств и занимают прочные позиции в инструментарии КУДО, мониторинга реализации образовательных программ, аудита качества образования. Возрастание объемов компьютерного тестирования, обусловленное расширенным внедрением дистанционных образовательных технологий в реализацию программ основного и дополнительного профессионального образования, не только повысило востребованность КТОЗУО и их роль в системе дидактических средств, но и существенно увеличило возможности критического осмысления и конструктивного анализа эмпирических результатов КУДО на различных этапах обучения на основе детерминированных и вероятностных математических моделей.

Несмотря на безусловный прогресс в совершенствовании методик разработки тестов и технологий их применения, КТОЗУО, как правило, играют вспомогательную роль в системе КУДО, выполняя значительный объем черновой работы при формировании оценок предварительной и промежуточной аттестации, контроля текущей успеваемости и остаточных знаний, а окончательное решение принимается посредством методик индивидуального опроса, беседы, письменных контрольных работ. Подтверждением этому отчасти является сокращение объема и значимости ТЗ с выбором единственного верного ответа в материалах и алгоритмах формирования оценок единого государственного экзамена, их частичная замена заданиями, предполагающими демонстрацию хода их решения или развернутый ответ.

Расширенная реализация дополнительных образовательных программ в дистанционном формате, стабильно высокий спрос на получение высшего образования и профессиональную переподготовку в заочной форме активизировали совершенствование методик дистанционного обучения, внедрение в учебный процесс специализированных программных продуктов, в том числе предназначенных для проведения занятий в оффлайн режиме и предполагающих оценивание

знаний и умений обучающихся. Сложности организации и проведения таких занятий во многом обусловлены противоречивостью дидактических требований к их структуре, содержанию и результатам, в частности необходимостью:

предварительного самостоятельного изучения обучающимися относительно больших объемов теоретического материала, его реферирования при ограниченности содержания контрольных вопросов, служащих для обучающихся ориентиром в выборе приоритетов изучаемого содержания;

совмещения обучающего и контролирующего компонента при ограничении времени доступа обучающегося к образовательному ресурсу;

унификации и оптимизации структуры оффлайн-занятий для снижения трудоемкости их разработки и переработки, обеспечения возможности многократного использования отдельных занятий в реализации различных образовательных программ;

соблюдения единства требований и критериев оценивания знаний обучающихся по различным учебным дисциплинам, обеспечения объективности оценок.

Такие организационные и технологические особенности КУДО в дистанционном формате, как невозможность или затрудненность оперативного обсуждения и критического анализа ответов обучающегося с преподавателем, апелляции, сочетания консультирования с контролем знаний и других форм диалога, обеспечивающих объективность оценки и ее осознание обучающимися, представляют собой фактор снижения действенности и оперативности обратных связей в учебном процессе, одновременно повышая значимость формирования или выбора комплекта контрольно-измерительных материалов в содержательном и процедурном аспектах.

Формирование ФОС образовательных программ опирается на многообразие контрольно-измерительных материалов, преимущественный выбор которых определяется задачами контроля, технологиями реализации образовательных программ и особенностями содержания обучения, степень освоения которого подлежит оцениванию. Актуальной тенденцией формирования ФОС является уменьшение доли ТЗ и возрастание доли заданий, требующих развернутого ответа, подробного аналитического решения, обоснованного разрешения ситуации и др. В последние несколько лет эта тенденция во многом обусловлена многочисленными фактами применения испытуемыми при ответах средств телекоммуникаций, в том числе в микро-

миниатюрном исполнении, обращения при выполнении ТЗ к интернет-ресурсам, включая чаты искусственного интеллекта (ИИ). В данных условиях не гарантируют достоверных оценок знаний и трудоемкие в проверке задания, требующие развернутого ответа, аналитического или числового решения, так как время, необходимое для их решения, оказывается достаточным и для использования испытуемым технологических возможностей. Относительно большое количество времени для решения такого рода заданий, наряду с ограниченными возможностями визуального контроля, является лимитирующим фактором их применения в дистанционных контрольных мероприятиях и испытаниях.

Неизбежное следствие уменьшение доли ТЗ в ФОС – возрастание трудоемкости контрольных мероприятий и снижение возможностей автоматизированной обработки результатов КУДО. Разрешение этого противоречия может быть достигнуто за счет применения и совершенствования ТЗ сложной структуры, диагностический и дидактический потенциал которых не уступает заданиям открытого типа, а применимость в автоматизированных системах мониторинга качества реализации образовательных программ более высокая.

Принципы построения тестов, отбора содержания, формирования алгоритмов оценивания и интерпретации результатов тестирования, обеспечивающие достижение наилучших показателей характеристик тестов, достаточно полно и всесторонне исследованы. Тем не менее в практике оценивания знаний обучающихся с применением тестов нередко возникают инциденты, ставящие под сомнение как объективность оценок, полученных испытуемыми, так и их верифицируемость другими способами оценивания знаний и умений контролируемого контингента обучающихся.

В условиях меняющихся подходов к выбору и реализации технологий КУДО особую актуальность приобретает проблема оценивания качества КТОЗУО через систему измеряемых показателей, способных характеризовать степень достоверности оценок учебных достижений. Дидактические тесты, применяемые для оценивания учебных достижений, характеризуются комплексом свойств, порой противоречивых, по-разному ранжируемых и трактуемых различными разработчиками тестов и тестологами и, как правило, не реализованных в адекватных математических моделях, что существенно затрудняет создание модели оценки качества теста. Такая модель, включающая в себя комплекс оцениваемых параметров, формализованные способы их оценки и анализа, критерии достижения необходимого качества или

оптимальности, должна согласовываться с моделями устройств и процессов, сходных по принципам и механизмам функционирования или реализации с КУДО.

Используемые многими исследователями методики оценки достоверности результатов тестирования, основанные на статистической модели Раша, оперирующей такими показателями, как трудность теста и вероятность его решения испытуемым с известным уровнем подготовки, опираются на обширную эмпирическую базу или оцененные ранее показатели. Данный подход, как правило, нереализуем в рамках относительно небольших по масштабам задач, включая разработку преподавателем вуза КТОЗУО для текущего контроля знаний, контроля внеаудиторной работы обучающихся, промежуточной аттестации и контроля остаточных знаний по преподаваемой учебной дисциплине, требующих оперативной и желательно априорной оценки качества компьютерных тестов.

Задача априорной оценки качества КТОЗУО в значительной мере упрощается с учетом применяемых для оценивания учебных достижений обучающихся двухбалльных («зачет – незачет») или четырехбалльных шкал, что позволяет отождествить КУДО с работой в первом случае отбраковочного автомата, а во втором – сортировочного автомата и, основываясь на этом, в качестве определяющей характеристики качества тестов выбрать надежность, в свою очередь выражающуюся в таких поддающихся измерению и априорным оценкам показателях, как безотказность, долговечность, ремонтпригодность и сохраняемость. В терминах теории надежности отказом теста будет считаться неверное оценивание знаний испытуемого. Относительно низкая прецизионность оценочных шкал, применяемых преподавателями для КУДО в системах высшего и среднего образования, позволяет трактовать дифференцирующую способность разрабатываемых ими КТОЗУО не как обеспечение множественности оценок, а как достижение высокой степени надежности разграничения диапазонов баллов, соответствующих принятой оценочной шкале. Потенциальная надежность такого разграничения, обеспечивающего достоверность оценок КУДО, повышается по мере уменьшения количества диапазонов тестовых баллов, т. е. оценок шкалы.

Данный подход может быть распространен на каждое ТЗ в отдельности, и тогда показатели надежности КТОЗУО (например, вероятность отказа или среднее время наработки на отказ) определяются как характеристики многокомпонентной системы с известными показателями надежности ее структурных элементов.

Создание таких моделей необходимо для корректного анализа различных видов ТЗ с целью выявления среди них обладающих относительно высокой потенциальной надежностью за счет структуры заданий и возможности их оперативной (с минимальными трудозатратами) переработки, направленной на восстановление контрольных функций теста, частично утрачиваемых вследствие его многократного использования или «взлома». Применение потенциально высоконадежных ТЗ, по нашему мнению, обладает большей перспективой повышения достоверности оценивания учебных достижений тестами и регулирования сложности тестов по сравнению с увеличением объемов баз заданий и выборок ТЗ с применением трудноверифицируемых эвристических и трудоемких эмпирических критериев оценивания результатов выполнения тестов.

Для решения задачи повышения достоверности тестирования учебных достижений в монографии последовательно рассматриваются общие подходы к оцениванию учебных достижений посредством тестов и проектированию КТОЗУО, дидактические характеристики различных видов ТЗ, модели априорной оценки достоверности результатов тестирования учебных достижений и качества дидактического теста, общий подход к моделированию его надежности и особенности учета видов ТЗ в моделях надежности тестов; исследуются некоторые способы повышения надежности КТОЗУО и достоверности результатов КУДО, включая корректность целеполагания КУДО, применение высоконадежных ТЗ, совершенствование алгоритмов начисления баллов за решение тестовых заданий и оценочных шкал; формируются практические рекомендации по разработке высоконадежных дидактических тестов.

1. ТЕСТ КАК ИНСТРУМЕНТ ОЦЕНИВАНИЯ УЧЕБНЫХ ДОСТИЖЕНИЙ

1.1. Технологии тестирования учебных достижений в современном образовании

Исторически значимым фактором деградации образования является плохая организация системы мониторинга учебного процесса, объединяющей в себе все формы контроля усвоения обучающими содержания образовательных программ, проявляющейся, с одной стороны, в тотальном либерализме преподавателей, с другой – в их субъективизме и предвзятости, порождающих множественные ситуативные критерии оценивания учебных достижений обучающихся.

Вошедшая в отечественную образовательную практику с начала 90-х гг. прошлого столетия система тестирования знаний и умений обучающихся, интерпретируемых как их учебные достижения (результаты обучения, готовность к освоению нового содержания обучения или новых образовательных программ, готовность к профессиональной деятельности), стала ведущим направлением модернизации контрольно-оценочных процедур, обеспечения единства требований на разных этапах обучения, мониторинга освоения образовательных программ, оценки эффективности образовательных систем и их структурных элементов, повышения качества обучения, обеспечения доступности профессионального образования [1].

К концу 80-х гг. XX в. в отечественном образовании утвердился технологический подход к построению моделей обучения, отождествляющий обучение или образование с технологическим процессом. На основе обобщения различных дефиниций понятия «технология» применительно к различным сферам жизнедеятельности и общественного производства мы рассматриваем технологию как *совокупность методов и средств, реализованных в процессах, операциях, процедурах, приемах и иных формах воздействия на объект, обладающий определенными исходными свойствами, с целью придания ему необходимых новых свойств.*

Объектом воздействия в образовательных технологиях являются обучающиеся, обладающие достаточным для обучения комплексом знаний и умений. В результате применения к ним необходимых средств и методов обучения, положенных в основу обучающих методик, которые, в свою очередь, реализуются в видах и формах проведения занятий, методических приемах и др., достигается требуемый

уровень обученности или квалификации, характеризующийся набором показателей, оцениваемых формализованными процедурами относительно просто (знаний, умений, навыков) или относительно сложно (владений, компетенций). Технологический подход к построению моделей обучения, по нашему мнению, является продуктом эволюции программированного обучения [2, 109], дидактические основы которого в отечественной педагогике активно разрабатывались в 60-е гг. XX в. и сопровождались внедрением методов кибернетики в управление образованием. Кризис программированного обучения в педагогической практике в 80-е гг. прошлого столетия во многом был обусловлен формальным применением законов кибернетики в проектировании процесса обучения, прогнозировании его результатов, управлении качеством и эффективностью на основе различных оптимальных критериев без учета психологических и социальных факторов, а также постепенной подменой понятия «программированное обучение» термином «обучение с применением компьютерных программ». Воспроизводя многие положения программного обучения, технологический подход к построению моделей обучения становится более гибким за счет их ориентации на расширенный и более детальный учет условий осуществления учебного процесса, использования численных методов и методов теории вероятностей в математическом моделировании. В современных моделях обучения, теоретическими основоположниками которых в 70–80-е гг. XX в. были В.П. Беспалько, Н.Ф. Талызина, Т.А. Ильина, Г.К. Селевко, М.В. Кларин и другие ученые [2, 3, 109–111], сам учебный процесс является объектом управления, подвергаясь циклическим или аperiodическим корректирующим управляющим воздействиям на основе обратных связей, что также имеет прямую аналогию с технологическими процессами.

В любых технологиях имеет место влияние случайных или сложно учитываемых факторов, которые не позволяют сделать конечный результат технологического процесса абсолютно детерминированным. Это выражается как в допустимых отклонениях от идеальных параметров конечного продукта, так и в допустимом проценте брака. Таким образом, любые технологии ориентированы на попадание требуемых характеристик конечного продукта в определенные интервалы (допуски) в окрестностях требуемых (расчетных, прогнозируемых) значений. Объективным условием снижения процента брака и уменьшения допусков характеристик конечного продукта является автоматизация технологических процессов, т. е. исключение из них или минимизация участия в них человека. В то же время несовершен-

ство автоматизированных компонентов технологического процесса, особенно на начальных этапах автоматизации, требует участия человека в доводке конечного продукта до кондиций неавтоматизированными манипуляциями, в том числе по корректировке параметров объекта на различных этапах технологического процесса на основе данных промежуточного контроля.

Любые технологии обеспечиваются множественными измерениями, направленными на контроль параметров как объекта на различных этапах технологического воздействия на него, включая входной контроль качества, так и самих процессов, т. е. соответствия текущих условий реализации технологий установленным требованиям. К измерениям предъявляются требования точности, характеризующейся допустимыми относительными погрешностями измерений.

Необходимо отметить следующие отличия образовательного процесса от производственных процессов:

- невозможность приведения исходных характеристик обученности контингента к единому уровню, обуславливающая сильный разброс индивидуальных показателей (знаний, способностей, личностных качеств) обучающихся в начале обучения;

- влияние на результаты обучения не только базовых знаний и интеллектуальных способностей обучающихся, но и их личностных качеств;

- большая зависимость результатов обучения от мотивации обучающихся, которая, в свою очередь, подвержена влиянию многих трудноучитываемых факторов, в том числе индивидуальных психологических реакций обучающихся;

- невозможность создания абсолютно одинаковых условий реализации образовательных технологий в различных местах их применения прежде всего из-за различий в квалификации и личностных особенностях профессорско-преподавательского состава образовательных организаций;

- отсутствие полной унификации (допустимость вариативности) объемов и содержания обучения, планов реализации одинаковых образовательных программ в различных вузах, их материально-технического и информационного обеспечения, методик обучения и воспитания;

- отсутствие унифицированных и воплощенных требований к организации, методам и формам контроля освоения обучающимися образовательных программ, особенно в межсессионный период, разработке и содержанию контрольно-измерительных материалов, монито-

рингу и управлению качеством образования в отдельных образовательных организациях.

Вышеизложенное обстоятельство приводит к существенным различиям в результатах обучения, что выражается в количественных показателях (оценках по балльной шкале), характеризующих достижение тех или иных маркеров освоения обучающимися образовательных программ или отдельных ключевых компетенций.

В связи с этим образовательные технологии характеризуются гибкостью, широкой вариативностью инструментария и набора обеспечивающих ресурсов, большой степенью индивидуализации применения, что во многом сближает их с медицинскими технологиями, в особенности с учетом социальной значимости результатов их применения.

Современные технологические процессы в значительной мере стандартизированы – унифицированные и нормированные требования охватывают как характеристики исходных и конечных продуктов, применяемых материальных средств, квалификацию специалистов, так и средства и методы контроля исполнения этих требований. По сути, исполнение стандарта начинается с условий контроля средствами измерений его требований, выраженных количественными параметрами, включая перечень таких параметров, их значения и пределы отклонения от них, средства измерений и методики измерений.

В настоящее время образовательные стандарты – федеральные государственные образовательные стандарты высшего образования, определяя необходимые условия реализации той или иной образовательной программы (например, для ее лицензирования или аккредитации), предъявляют единые требования к качеству подготовки специалистов лишь в формате перечня части компетенций, которые должен освоить выпускник.

Дополнение списка компетенций профессиональными компетенциями, их детализацию (декомпозицию), формирование программ государственной аттестации осуществляет образовательная организация или ведомственный орган, организующий подготовку специалистов в интересах ведомства. Требования к организации контроля качества обучения в действующих стандартах изложены безотносительно критериев, норм, объемов, регулярности, способов и средств оценивания. Иными словами, в отличие от технических стандартов, образовательные стандарты не регламентируют количественные параметры, характеризующие достижение результатов применения образовательных технологий, способы и средства их оценивания. Данное об-

стоятельство обязывает образовательные организации самостоятельно формировать системы контроля качества реализации образовательных программ, в том числе средства и методы контроля.

Предполагается, что переход в 2026–2027 учебном году на новую модель высшего образования будет сопровождаться более определенными требованиями к оценке качества подготовки специалистов как в содержательном, так и в организационном и технологическом аспектах. Данное обстоятельство призвано подчеркнуть значимость КУДО в реализации образовательных программ, активировать и обобщить передовой опыт вузов в освоении и развитии инструментов и технологий оценивания знаний и умений обучающихся на различных этапах обучения.

Именно технологический характер образовательного процесса обуславливает поэтапный КУДО как инструмент управления качеством образования. Для любой технологии важно регулярное измерение текущих характеристик объекта с целью прогнозирования качества готового продукта и при необходимости внесения корректировок в параметры технологических процессов. Такие измерения представляют собой начальный элемент цепи обратной связи. Они обеспечивают управление системой, а результаты измерений являются исходными данными для выработки управленческих решений (формирования управляющих сигналов).

Еще в середине 70-х гг. XX в. Н.Ф. Талызина отмечала необходимость циклического управления процессом обучения по принципу «белого ящика», т. е. с учетом выявленных закономерностей и структурных взаимосвязей, которые следует использовать для формирования цепи обратной связи. Коррекцию процесса усвоения содержания обучения за счет действия обратной связи исследователь рассматривала как самостоятельный и важный этап управления, а цепь обратной связи – как фактор завершенности системы управления. В свою очередь, такая коррекция осуществима только при возможности получения достоверной, объективной информации о состоянии образовательной системы.

Несмотря на методологическое разнообразие современной теории и практики управления качеством обучения, общими положениями актуальных моделей управления учебным процессом являются:

- 1) управление учебным процессом (один из определяющих факторов повышения его эффективности);
- 2) актуальная и достоверная информация о показателях качества обучения (необходимое условие управления учебным процессом).

Таким образом, достижение высокого качества обучения возможно только при использовании объективных методов оценивания его показателей. Эволюция методов диагностики учебных достижений обучающихся весьма показательна.

Первый этап расширенного внедрения тестов для оценивания учебных достижений, которые для отграничения от психологических тестов ряд специалистов называет педагогическими тестами [4] или, подчеркивая применение в учебном процессе, дидактическими тестами [5], в отечественной педагогике был следствием разработок и реализации моделей программированного обучения, т. е. КУДО тестами был элементом технологии программного обучения. На фоне критики субъективности КУДО в традиционных, устных формах ТОЗУО стали внедряться в различные контрольно-оценочные процедуры. Большой объем эмпирических данных позволил избрать гибкую тактику применения тестов в качестве факультативных, альтернативных, основных оценочных процедур или в сочетании с традиционными формами оценивания.

Оценивание учебных достижений обучающихся относят к педагогическим измерениям, которые определяются как «процесс установления соответствия между оцениваемыми характеристиками обучающихся и точками эмпирической шкалы, в которой отношения между различными оценками характеристик выражены свойствами числового ряда» [6]. Помимо определения количественных параметров учебных достижений, педагогические измерения могут иметь более широкое предназначение (например, выявление мотивации к обучению, лидеров в учебе, степени соответствия условий обучения образовательным целям).

Большой вклад в развитие теории и практики педагогических измерений внесли исследования фундаментального характера и публикации Д. Батесона, А. Бирнбаума, Г. Раша, Н.М. Розенберга, В.С. Аванесова, В.П. Беспалько, А.Н. Майорова, В.И. Нардюжева, А.П. Попова, А.О. Татура, В.А. Хлебникова, М.Б. Челышковой, А.Г. Шмелева и др.

Одним из направлений критики традиционных форм КУДО является малая прецизионность используемых оценочных шкал – пятибалльной или четырехбалльной. Так, буквальное применение правил исчисления погрешностей приводит к тому, что при использовании пятибалльной шкалы преподаватель выставляет оценки с минимальной относительной погрешностью 20%, а за один и тот же ответ испытуемого его знания могут быть оценены разными экзаменаторами

на «2», «3» и «4» [7]. Основываясь на такой интерпретации положений теории и техники измерений физических величин, легко прийти к выводу, что оценивание по шкале «зачтено – не зачтено» имеет относительную погрешность 50%, а испытуемый может равновероятно получить зачет или незачет.

В данном случае имеет место неверное отождествление точности измерений с прецизионностью измерительной шкалы. Высокая прецизионность шкалы средства измерений, т. е. способность различать близкие значения измеряемой величины, является одним из необходимых, но далеко не достаточным условием его точности. Так, в начале XX в. первичные (наиболее точные) эталоны метра, стержни X-образного сечения из иридиевого-платинового сплава, хранившиеся в лабораториях наиболее технически развитых стран (США, Германии, Великобритании, России, Франции) и являвшиеся копиями международного эталона, воспроизводили метр с максимально достижимой на тот момент погрешностью в 1 мкм (10^{-6} м). Обладая каждой шкалой из всего 10 делений, они должны характеризоваться низкой прецизионностью, однако точность воспроизведения ими длин 0,1; 0,2; 0,3... 1,0 м и в настоящее время существенно превышает точность намного более прецизионных измерительных линеек, рулеток и др.

Использование низкопрецизионных четырех- и пятибалльных шкал на практике означает соотнесение оцениваемых экзаменатором знаний и умений обучающихся с некоторым интервалом на оценочной шкале. В результате испытуемому выставляется оценка, являющаяся нижней границей данного интервала. Данный механизм оценивания обладает высокой потенциальной надежностью, а низкая объективность КУДО в устной форме, как показывает практика, обусловлена отсутствием четких критериев оценок или их несоблюдением экзаменатором. Надежность оценивания знаний и умений обучающихся человеком тем выше, чем ниже прецизионность оценочной шкалы – самой потенциально высокой надежностью обладает оценивание по шкале «зачтено – не зачтено», и напротив, практика использования шести-, десяти- и двадцатибалльных шкал для КУДО выявляет снижение воспроизводимости оценок при повторном контроле и повышение вариативности оценок, выставляемых разными педагогами. Параллельное использование низкопрецизионных и высокопрецизионных оценочных шкал создает проблемы в их взаимном согласовании: если, например, перевод из стобалльной шкалы осуществляется относительно однозначно (хотя, как показывает практика ЕГЭ с применением модели Раша [4, 8, 9], пороговые значения оценок пяти-

балльной шкалы являются аргументами нелинейной функции и могут быть не определены априори), то обратный перевод либо методически некорректен, либо, будучи осуществленным, обеспечивает лишь частичное сопряжение двух шкал (например, при суммировании полученных по ним оценок). Вследствие этого результаты ТОЗУО достаточно долгое время приводились к привычной для восприятия пятибалльной (четырехбалльной шкале), а способность дидактических тестов оперировать так называемыми интервальными оценочными шкалами, позволяющими применять для их обработки методы математической статистики, практически не использовалась.

На практике вопросы валидности тестов, т. е. их пригодности для КУДО, в обоих аспектах – содержательном (то ли содержание обучения оцениваем) и измерительном (обеспечивает ли тест достоверную оценку) – решались разработчиками тестов самостоятельно на основе эмпирико-эвристических подходов. Сам термин «валидность теста» привнесен в дидактическое тестирование из психологического тестирования (психологической диагностики), где обеспечение валидности теста является зачастую труднорешаемой задачей, так как, согласно принятой терминологии, психологическими тестами оцениваются латентные (т. е. скрытые, выраженные неявно, опосредованно, через прямоизмеряемые параметры) характеристики. Согласно теории измерений такие измерения являются косвенными, их результат выражается через результаты прямых измерений, которые, в свою очередь, непосредственно отображаются измерительной шкалой. Знания и умения обучающихся оцениваются дидактическими тестами напрямую, непосредственно, т. е. численное значение (балл оценочной шкалы) присваивается испытуемому исходя из результатов решения им ТЗ без измерения каких-либо вспомогательных или промежуточных параметров: каждому ТЗ априори ставится в соответствие балл за его решение (или различные баллы за частично верные решения) в зависимости от трудности, сумма баллов за правильно и частично правильно решенные ТЗ (возможно, за вычетом «штрафов» за некоторые нерешенные ТЗ) и образует итоговую оценку. По нашему мнению, будучи прямоизмеряемыми, знания и умения обучающихся не могут быть отнесены к латентным переменным, в отличие, например, от когнитивных способностей, характеризующих обучаемость, – памяти, восприятия, логического мышления и др., количественные оценки которых могут быть получены в ходе неоднократных применений дидактических тестов. Следовательно, и проблема обеспечения валидности ТОЗУО не является самостоятельным предметом теоретическо-

го рассмотрения. Проблемы с содержательной валидностью обусловлены исключительно человеческим фактором, т. е. низкой квалификацией разработчика теста, и, согласно теории измерений, могут решаться двумя путями: повышением до достаточного уровня квалификации разработчика теста либо исключением человека из процесса разработки тестов. В свою очередь, измерительный аспект должен рассматриваться исключительно в рамках метрологии и квалиметрии, располагающими исчерпывающим набором стандартизированных решений, в том числе на основе современных математических моделей.

В работе Н.Ф. Талызиной приведен пример неудачного использования контролирующих устройств (тестеров), программа контроля (тест) для которых разрабатывалась различными преподавателями [3]. Последовательная проверка одного и того же контингента учащихся по одной и той же теме, но по разным тестам показала различные уровни достижений, из чего сделан вывод о невалидности тестов и невозможности их использования в КУДО из-за низкой объективности контроля. Данный пример демонстрирует выявившуюся при расширенном внедрении ТОЗУО проблему квалификации разработчиков дидактических тестов и в целом влияния человеческого фактора на объективность КУДО, а также априорного оценивания надежности тестов и достоверности получаемых по ним результатов. В частности, один из применяемых тестов может обеспечивать достоверность оценок, т. е. объективность контроля, соответственно, остальные – невалидны. Признание всех дидактических тестов невалидными основывается лишь на несовпадении оценок знаний испытуемых, полученных с применением каждого теста, так как проблемы с сопряжением оценочных шкал и сопоставлением оценок знаний, полученных различными способами, не позволили осуществить независимую верификацию тестов.

Не менее сложной задачей для разработчика психологического теста является достижение высоких показателей его надежности, выражающейся в воспроизведении результатов при повторных тестированиях. Будучи перенесенной на разработку дидактических тестов, проблема обеспечения высокой надежности на практике вызвала трудности, связанные с повторяемостью индивидуального количества баллов при многократном применении теста. Корректность целеполагания в данном случае ставится под сомнение в связи со следующим:

одинаковая сумма баллов может быть набрана большим количеством способов;

при каждом повторном тестировании в тестовой выборке неизбежно наличие ранее решенных ТЗ, и чем таких ТЗ больше, тем выше будет итоговый балл повторного тестирования;

чем больше новых ТЗ используется в повторном тестировании, тем больше отличается новый тест от предыдущего, что делает постановку вопроса о надежности теста неуместной.

Наконец, многократное повторение неверного результата оценивания учебных достижений не может свидетельствовать о надежности ТОЗУО, так как в данном случае тест не выполняет своей функции, что в теории надежности интерпретируется как отказ.

Следующий этап расширенного внедрения ТОЗУО обусловлен компьютеризацией учебного процесса в конце 80-х – середине 90-х гг., что сделало доступным для обработки и интерпретации результатов тестирования специализированное программное обеспечение, а также существенно упростило разработку и применение тестирующих программ и в дальнейшем обеспечило возможность размещения тестов в специализированных программных приложениях – тестовых оболочках. Компьютеризация КУДО, выполняемого посредством дидактических тестов, позволила реализовать одно из главных преимуществ тестирования перед устным оцениванием знаний – производительность, а также обеспечила полное исключение человека из процесса проверки и оценивания. Развитие сетевых технологий в середине – конце 90-х гг. дало возможность управлять КУДО на компьютерах одновременно в нескольких компьютерных классах, оперативно аккумулировать и анализировать результаты КУДО большого количества испытуемых, что создало предпосылки для разработки и внедрения в образовательных организациях систем управления качеством обучения, основанных на независимом внутреннем и внешнем контроле знаний и умений обучающихся на различных этапах освоения ими содержания обучения, а также внешнего аудита качества обучения со стороны уполномоченных контрольных органов. Оперативное получение больших массивов оценочных данных, в том числе при повторных КУДО, позволило создать и апробировать корректные статистические модели, оперирующие высокопрецизионными оценочными шкалами, что создало предпосылки для внедрения в России технологий ЕГЭ и обеспечило прозрачность для общества результатов деятельности всей отечественной системы общего образования.

Компьютеризация процедур КУДО постепенно, по мере освоения нового инструментария реализации дидактических тестов, позволила существенно уменьшить влияние на оценки испытуемых так

называемого «парадокса артефакта», при котором измеряется не истинное свойство объекта, а результат проявления этого свойства при воздействии на объект измерительного инструмента. Критики оценивания знаний в форме собеседования приводят множество примеров, когда преподаватель вольно или невольно оказывает влияние на оценку, производя на испытуемого впечатление, что он его «валит», «давит психологически», «запутывает», «расслабляет, не дает сосредоточиться», «провоцирует на дачу неверного ответа», «сам говорит верный ответ, когда его уже вот-вот готов сформулировать отвечающий», «отвлекает», «уводит в сторону своими рассуждениями» и т. п. Однако и дидактический тест как измерительный инструмент оценивает не сами знания испытуемого, а их проявление в конкретных условиях в установленный момент времени именно вследствие воздействия ТЗ на память и мышление испытуемого. Следовательно, и дидактический тест вносит в измерения некоторую систематическую (методическую, инструментальную) погрешность, обусловленную как несовершенством самого теста (например, неполнотой проверяемого содержания обучения), так и влиянием на результаты тестирования психологических реакций испытуемого. Будучи относительно редко употребляемым инструментом оценивания, дидактические тесты и условия их применения (письменное, а иногда и невербальное решение ТЗ влечет риски, связанные с описками, ошибками из-за невнимательности, подразумевает жесткий лимит времени, невозможность исправить замеченную ошибку или сильное влияние такого исправления на итоговый балл и т. п.) у некоторых испытуемых вызывают стрессовые реакции, а отсутствие при ответе обратной связи с преподавателем и, как следствие, существенно более низкий по сравнению с устным контролем знаний дидактический потенциал ТОУЗО порождают негативное восприятие данной формы контроля как антигуманной («проверочный конвейер»). Зачастую чем более харизматичны преподаватели, читающие лекции и ведущие занятия семинарского типа, тем они сильнее. Однако по мере приобретения опыта решения ТОЗУО, совершенствования интерфейса тестовых программных оболочек влияние дидактического теста как измерительного инструмента на результат его применения при КУДО сводится к минимуму. В период эксперимента по проведению ЕГЭ в регионах России в 2001–2008 гг. многочисленные ошибки испытуемых при заполнении проверочных бланков привели к необходимости проведения с учащимися выпускных классов не только многократных инструктажей,

но и «пробных» ЕГЭ, что позволило снизить в несколько раз процент ошибок, связанных с работой с бланками.

Дальнейшему развитию КУДО на основе дидактических тестов послужило расширение применения ДОТ в конце 90-х гг. и первой половине первого десятилетия XXI в. на базе технологий удаленного доступа к образовательным ресурсам. ДОТ в высшем образовании, особенно в реализации программ переподготовки и повышения квалификации, заняли обширную нишу, которая до настоящего времени не уменьшилась, в том числе благодаря развитию технологий дистанционного КУДО в онлайн-режиме. Обеспечение объективности и достоверности оценок испытуемых в формате компьютерного тестирования при проведении промежуточной и итоговой аттестации при реализации образовательных программ дополнительного профессионального образования стало ключевым моментом внедрения ДОТ в данный образовательный сегмент и позволило за счет сокращения временных и ресурсных затрат на обучение максимально удовлетворить потребности в повышении квалификации и профессиональной переподготовке как работников, так и их работодателей.

Современные специализированные программные среды, такие как CVO MOODLE (от англ. Modular Object-Oriented Dynamic Learning Environment), обеспечивают разработчикам дидактических тестов широкие возможности в части использования в ТЗ графических и видеоматериалов, видового разнообразия применяемых ТЗ, управления различными параметрами тестирования – объемом выборки из базы ТЗ, лимитом времени решения теста и отдельных ТЗ, тактики решения теста (с возможностью пропуска отдельных ТЗ и последующего возврата к ним), а также реализации обучающих возможностей теста – демонстрации верных ответов по окончании тестирования, работы над ошибками с повторным контролем и др. Это способствовало распространению опыта дистанционного КУДО на образовательные программы высшего образования, осваиваемые по заочной форме.

Проверка остаточных знаний обучающихся при проведении аккредитации образовательных организаций высшего и среднего профессионального образования с применением дисциплинарных компьютерных тестов позволила унифицировать контрольно-аттестационные процедуры и создать основу для объективного сравнительного анализа результативности образовательной деятельности вузов и их учебных подразделений.

Повышение востребованности КТОЗУО в последние годы обусловлено в том числе необходимостью достоверного определения уровня освоения обучающимися компетенций, предусмотренных федеральными государственными образовательными стандартами или квалификационными требованиями. Поскольку формирование различных компетенций завершается не одновременно и условия для их поддержания на актуальном уровне до выпуска специалистов объективно не равны, большой интерес как для образовательной организации, так и для работодателей ее выпускников представляет оценка уровня освоения компетенций обучающимися на завершающем этапе обучения, непосредственно перед государственной итоговой аттестацией и в процессе нее. Ввиду большого количества компетенций, предусмотренных основными образовательными программами, а также индикаторов их освоения далеко не все они могут быть оценены в ходе государственного экзамена или защиты выпускной квалификационной работы, что приводит к необходимости применения для оценивания уровня сформированности компетенций дополнительных оперативных (из-за ограниченных сроков проведения) оценочных процедур [10–14].

Прогресс в телекоммуникационных технологиях, способствовавший очередному этапу расширенного применения в учебном процессе ТОЗУО, создал и дополнительные трудности КУДО, обусловленные несанкционированным применением испытуемыми телекоммуникационных средств. Смартфоны, смарт-часы, микроминиатюрные передатчики, беспроводные микрогарнитуры, в том числе скрытно размещаемые внутри ушной раковины, практически обесценили в дидактическом аспекте офлайн-тестирование из-за невозможности или технически неоправданной сложности обеспечения условий объективного оценивания знаний испытуемых. Дидактические тесты для контрольных процедур в онлайн-формате в условиях противодействия угрозам объективности оценивания знаний и умений обучающихся претерпели структурную модернизацию, направленную на нейтрализацию скрытого оперативного доступа недобросовестных испытуемых к неразрешенным источникам информации. Все более значительную часть структуры таких тестов занимают ТЗ открытого типа, предполагающие краткий или в меру развернутый ответ на прямой вопрос, и ТЗ с усложненной структурой, ответ на которые требует не просто формального знания, легко воспроизводимого подсказкой, но и осмысления контролируемой ДЕСО и связи ее содержания с содержанием смежных дидактических единиц. Разработчики ТОЗУО

все чаще стали включать в базы ТЗ вопросы-ловушки, направленные на перепроверку уже данных ответов или косвенную верификацию испытуемого. Таким образом, квалификация современного разработчика ТОЗУО должна обеспечивать проектирование тестов, удовлетворяющих формальным показателям качества и обладающих повышенной надежностью в условиях воздействия деструктивных факторов.

Современный этап эволюции системы контроля качества обучения посредством регулярного оценивания учебных достижений обучающихся характеризуется проникновением в нее ТИИ, что имеет двоякие последствия. С одной стороны, обращение к чатам ИИ позволило ускорить разработку ТЗ относительно простой структуры (с выбором единственного верного ответа, на установление соответствий) и роль разработчика ТЗ может быть сведена к редактированию предлагаемых чатом ИИ заготовок, значительная часть которых вполне может оказаться пригодной к применению без редактирования. С другой стороны, неконтролируемый и несанкционированный доступ к чатам ИИ испытуемых в ходе контрольных мероприятий позволяет им получать ответы на ТЗ в режиме реального времени с максимальной оперативностью. С учетом того что часть ТЗ может быть создана ИИ, чат ИИ будет тестировать сам себя или другой чат ИИ с трансляцией результатов решения ТЗ через испытуемого. Даже при собеседовании с экзаменатором, когда подключенный чат ИИ транслирует испытуемому на скрытую микрогарнитуру ответы на вопросы экзаменатора или зачитываемые испытуемым вопросы билета, у экзаменатора может сложиться убедительная иллюзорная картина осведомленности испытуемого в предъявленном к контролю содержании обучения. Данное обстоятельство требует дальнейшего повышения надежности ТОЗУО путем включения в них ТЗ усложненной структуры (с градацией верных ответов, с упорядочиванием дистракторов, на установление последовательности), а также замены вербальной информации графической (представление вопросов и вариантов ответов графиками, диаграммами, фотографиями и т. п.).

Наконец, применение ТИИ в анализе результатов тестирования призвано упростить и одновременно повысить надежность интерпретации числовых данных, исключить ошибки субъективной природы и сократить до минимума количество задач, решаемых на основе экспертных и эвристических методов.

Злоупотребление испытуемыми возможностями средств телекоммуникаций и ИИ повысило интерес педагогов практиков к интерактивным формам контроля, применяемым в ходе деловых игр, уче-

ний, решения ситуативных задач. Будучи основным оценочным средством в образовательных программах, реализуемых дистанционно, ТОЗУО при обучении в очном формате не играет самостоятельной роли как оценочное средство, а используется в специальных оценочных процедурах внутреннего и внешнего аудита качества обучения или совместно с другими формами текущего контроля и промежуточной аттестации. Характеризуясь высокой производительностью, эффективным аналитическим обеспечением, совместимостью с различными технологиями обучения, ТОЗУО обладают типичными достоинствами и недостатками обезличенного инструмента в педагогическом процессе. К последним стоит отнести довольно низкий обучающий потенциал, отсутствие собственной мотивирующей функции, подверженность деградации при многократном использовании и необходимость дополнительного контроля для обеспечения объективности результатов. Для сохранения своей роли в процессе обучения дидактическое тестирование должно эволюционировать в направлении повышения надежности ТОЗУО, что в условиях современного технологического обеспечения КУДО может быть реализовано главным образом в разработке и использовании высоконадежных ТЗ и совершенствовании структуры тестов.

1.2. Основные подходы к проектированию тестов для оценивания и контроля учебных достижений обучающихся

При оценивании учебных достижений обучающихся многие специалисты рассматривают дидактический тест как наиболее объективное средство измерения, обеспечивающее возможность массовой, оперативной, многомерной диагностики результатов учебной деятельности.

Тесты для оценивания знаний и умений обучающихся как объективное диагностическое средство предоставляют информацию для анализа процессов в отдельных образовательных организациях, образовательных системах региона или отрасли (ведомства), сфере образования в целом, которая характеризуется такими качествами, как точность, полнота, достаточность, системность, оптимальность, обобщенность, оперативность и доступность.

Актуальными преимуществами КТОЗУО перед другими формами контроля являются большой выбор программных средств для создания тестов, интегрированность КТОЗУО в программные оболочки

систем дистанционного обучения, оперативность контроля и анализа его результатов, сформированные научные основы разработки и огромный практический опыт применения тестов и ТЗ для различной целевой аудитории [1, 35, 36].

В [8] процесс тестирования учебных достижений имеет три этапа: 1) разработка теста; 2) процедура тестирования; 3) обработка и интерпретация результатов тестирования. По нашему мнению, в условиях массового применения теста важной процедурой является апробация теста или его экспертиза с возможной корректировкой, которая может быть включена в первый этап или представлять собой обособленный этап [40, 42, 87, 108].

Для понимания современных методов и технологий разработки и применения ТОЗУО необходимо рассмотреть основные понятия и термины, связанные с тестологией, в порядке, отражающем последовательность действий от целеполагания тестирования до обработки результатов КУДО.

Тестология (в переводе с англ. *test* – проба, греч. *logos* – знание) имеет несколько значений: с одной стороны, это наука об испытании, а с другой – наука о методах педагогических измерений и их интерпретации [15].

Терминология в сфере педагогических измерений характеризуется свойственной в целом для данной отрасли квалиметрии системной проблемой, порожденной несоответствием употребляемых терминов и понятий их аналогам в других отраслях знания. Такая терминологическая многозначность, а иногда и омонимичность, обусловлена либо некорректным заимствованием терминов из смежных предметных областей или конструированием их эквивалентов, либо разработкой и развитием педагогической квалиметрией собственной терминологической базы и понятийного аппарата без учета уже сложившихся веками и стандартизированных толкований терминов аналогичного написания и звучания.

Согласно [16] *измерение* – это процесс экспериментального получения одного или более значений величины, которые могут быть обоснованно приписаны величине. В [16] предельно обобщены вводимые ранее различными стандартами [17, 18] определения измерения, в частности исключено участие в измерении специальных технических средств, что расширило применение данного термина на нефизические объекты.

Наиболее близким к данному стандартизированному определению является приведенное в [19]: *измерение* – процесс установления

соответствия между некоторой совокупностью объектов (измеряемых параметров) и множеством чисел в соответствии с определенными правилами.

Анализ аналогий и сходств педагогических и физических измерений, терминов метрологии и педагогической квалиметрии позволяет выделить в измерении следующие этапы:

- 1) определение объектов измерения (измеряемых параметров);
- 2) определение целей измерения: диагностика, контроль параметров, инициативное исследование, обновление (пополнение, формирование) базы эмпирических данных и др.;
- 3) выбор методов измерения и измерительных процедур (методик);
- 4) выбор (конструирование, оптимизация) измерительных инструментов и их интегрирование в измерительные процедуры;
- 5) выбор (конструирование) измерительной шкалы;
- 6) получение массива экспериментальных данных и их формализация, т. е. присвоение им первичных числовых значений;
- 7) отображение результатов измерения на измерительную шкалу по определенным правилам (алгоритмам);
- 8) обработка и интерпретация результатов измерения.

В теории педагогических измерений существует много определений понятия *педагогического теста*, заметно отличающихся друг от друга.

Т.М. Балыхина рассматривает *педагогический тест* как комплекс заданий, измеряющих уровень учебных достижений, обученности, прогресс в учебной деятельности, эффективность учебного процесса [20]. Перечисленные объекты измерений очень разнородны в определении возможных измеряемых параметров. Очевидно, «уровень учебных достижений» и «прогресс в учебной деятельности» как комплекс статических величин и производных от него показателей характеризуются различными количественными показателями, способами получения эмпирических данных и алгоритмов их обработки. Еще в большей степени это относится к «эффективности учебного процесса», если учесть, что необходимым условием расчета эффективности является знание ресурсных затрат на получение искомого результата.

Согласно определению, *педагогический тест* – это квалиметрически выверенная система ТЗ, методов их предъявления и оценивания результатов их выполнения, которая обеспечивает получение наиболее обоснованных характеристик объекта испытания [21]. Данная дефиниция опирается на неоднозначно трактуемые формулировки,

в частности «наиболее обоснованные характеристики», «квалиметрически выверенная система», а также предполагает, помимо собственно оценивания результатов, еще и их интерпретацию, что по смыслу соответствует приведенным выше определениям измерения, инструментом которого является тест.

В.С. Аванесов предлагает следующую дефиницию: *педагогический тест* – система параллельных заданий возрастающей трудности, специфической формы, которая позволяет качественно и эффективно измерить уровень и структуру подготовленности испытуемых [9, 22]. В данном определении спорной является возрастающая трудность заданий, определяющая не столько дифференцирующую способность теста, сколько алгоритм тестирования «от простого к сложному», нечасто реализуемый в тестовых процедурах и отражающий скорее специфику адаптированного теста [23, 24]. Требование «параллельности», т. е. независимости, заданий далеко не всегда гарантирует объективность оценивания знаний и умений обучающихся, так как «связанные» задания позволяют выявить реальное знание соответствующей ДЕСО: если одно из них решено правильно, а другое – нет, это говорит о том, что правильный ответ был известен испытуемому заранее или был им угадан, т. е. в целом о незнании данной ДЕСО или недостижении измеряемого индикатора компетенции. Указание на специфическую форму заданий не несет полезной смысловой нагрузки ввиду многообразия таких форм, а задания открытого типа (например, используемых в разделах «b» и «с» тестов ЕГЭ) позволяют применить рассматриваемое определение к любой форме контроля знаний и умений, реализуемой письменно или на компьютере.

В.С. Ким полагает наиболее приемлемым следующее определение: «*Педагогический тест* – это система ТЗ различной трудности, которая позволяет качественно и эффективно измерить уровень и структуру подготовленности испытуемых» [8]. Отдавая должное лаконизму формулировки и соглашаясь с ее автором в части различной трудности ТЗ, не можем не отметить некорректность употребления понятия «подготовленность» в аспекте проверки результатов обучения. Обычно подготовленность соотносят с каким-либо видом деятельности, однако вряд ли тест может быть эффективно применен для измерения уровня подготовленности каменщика, газосварщика или оператора беспилотного воздушного судна, в значительной мере определяемого уровнем практических умений и навыков. Очевидно излишне употребление в данной дефиниции понятия эффективности применительно к педагогическому измерению, поскольку, обращаясь

к наиболее общему определению эффективности как к отношению результата к затратам на его достижение, мы вынуждены признать, что самым эффективным инструментом измерения подготовленности испытуемых является интуиция – она дает ненулевой (с вероятностью угадывания) результат при нулевых затратах. Что же касается качества измерений, то под ним в метрологии понимается целый комплекс регулируемых характеристик (достоверность, точность, надежность, сходимость и др.). Словосочетание «качественно измерить» трактуется как решение измерительных задач, где не требуется высокая точность измерений, а необходимо выявить качественные отличия исследуемых объектов. В данном контексте более уместно употребление выражения «измерить с необходимым (заданным) качеством».

А.Н. Майоров на основе обобщения множественных определений теста дает следующее: «Тест – это инструмент, состоящий из квалитетически выверенной системы ТЗ, стандартизированной процедуры проведения и заранее спроектированной технологии обработки и анализа результатов, предназначенный для измерения качеств и свойств личности, изменение которых возможно в процессе систематического обучения» [25]. Отметим, что в данном определении речь идет именно о педагогическом тесте, так как объект его применения выходит за рамки знаний и умений или результатов обучения в целом.

М.Б. Челышкова определяет *итоговый нормативно-ориентированный тест* как систему ТЗ, «упорядоченных в рамках определенной стратегии предъявления и обеспечивающих информативность оценок уровня и качества подготовки испытуемых» [6].

В своих рассуждениях мы исходим из того, что педагогическое измерение является более общим понятием по отношению к измерениям знаний и умений обучающихся. Следовательно, педагогический тест является более общим понятием по отношению к ТОЗУО.

Под ТОЗУО будем понимать измерительное средство, представляющее собой систему унифицированных калиброванных заданий и связанную с ней процедуру начисления баллов и их соотнесения с оценочной шкалой, позволяющую достоверно оценить уровень учебных достижений испытуемых [26].

В наших публикациях мы используем два близких по смыслу понятия – ТОЗУО и ТОУД. Разница состоит в том, что ТОУД является инструментом контроля учебного процесса, а ТОЗУО реализует как контрольные, так и диагностические функции.

Согласно определению ТОЗУО включает в себя три компонента:

1) ТЗ, систематизированные по какому-либо принципу (например, отобранные в соответствии с ДЕСО, распределенные по уровням

трудности, ориентированные на применение различных или однотипных видов ТЗ);

2) процедуру перевода ответов испытуемых в баллы или иные оценочные показатели – оценивания в баллах или иных подлежащих дальнейшей обработке и анализу информативных показателей (символах, цветах, отождествляющих эмоции «смайликах» или междометиях и др.) отдельных ТЗ в зависимости от их уровня трудности и вида;

3) процедуру перевода полученных баллов или иных информативных показателей в итоговую оценочную шкалу, однозначно характеризующую уровень проверяемых знаний и умений.

Систематизация заданий предполагает наличие некоторой *базы заданий* и процедуры отбора и компоновки заданий в тест. В общем случае количество заданий в базе N_B превышает количество заданий в тесте: $N_B > N_3$, однако на этапе отладки теста или при однократном применении теста приемлемо равенство N_B и N_3 .

Перевод ответов испытуемых в баллы для каждого ТЗ производится на основе сличения ответа с эталонным решением этого задания (эталонном), хранящимся в *базе знаний*. Совокупность эталонных решений для одного теста является его *ключом*.

Для обоснованного сопоставления результатов обучающихся между собой тестовые баллы, в соответствии с рядом критериев и норм, переводятся в производные показатели при помощи процедуры *шкалирования*. Преобразование шкал на основе анализа статистических результатов нормативной выборки позволяет выставить каждому испытуемому тестовый балл вне зависимости от того, в какой группе и над каким вариантом теста он работал [27].

Структура теста (распределение выбранного для оценивания количества ТЗ по видам и уровням трудности), его оценочная шкала, процедура перевода баллов в оценки определяется целями тестирования, в связи с чем дидактические тесты подразделяются на два типа – *нормативно-ориентированные* и *критериально-ориентированные* [4, 6]. Назначение нормативно-ориентированных тестов – прецизионная диагностика и ранжирование испытуемых, они применяются в модульно-рейтинговой системе оценивания, для проведения вступительных испытаний в образовательных организациях, в итоговых оценочных мероприятиях выпускников, в научных исследованиях. Такие тесты обладают высокой дифференцирующей способностью, высокопрецизионными шкалами, предполагают сложные, зачастую апостериорные процедуры перевода баллов за решенные ТЗ в итоговые оценки или сложные процедуры назначения баллов за решенные

ТЗ в зависимости от их трудности, вида и полного или частично верного решения. Высокая прецизионность оценочных шкал нормативно-ориентированных тестов создает предпосылки для широкого применения методов математической статистики для обработки результатов измерений, вплоть до построения функций распределения и плотностей распределения вероятностей больших массивов эмпирических данных с вычислением моментов статистических функций, определения детерминированных и случайных компонентов оценок, включая систематические и случайные компоненты погрешностей оценивания. Для небольших объемов тестирования практический интерес вызывает сравнение медиан и мод отдельных выборок [8, 26]. При нормативно-ориентированном тестировании первоочередной задачей является не столько выявление полноты освоения содержания обучения, сколько относительная дифференциация испытуемых по уровню подготовки. Рейтинг тестируемых представляет собой основу интерпретации результатов: для каждого испытуемого определяется процент тестируемых, у которых результат теста хуже или лучше его собственного [6]. Трудность заданий такого теста лежит в широком диапазоне – от самых простых до очень трудных. Примерами нормативно-ориентированной интерпретации результатов педагогических измерений являются тесты ЕГЭ и централизованного (абитуриентского) тестирования.

Критериально-ориентированные тесты предназначены для проверки по заданным критериям успешности усвоения определенного содержания обучения, наличия необходимых знаний и умений индивидуально и в группах. Они применяются в системах КУДО на различных этапах обучения испытуемых, аттестации выпускников, освоивших программы основного или дополнительного образования, для оценки достижения ими минимально допустимого уровня компетентности в системах внешнего и внутреннего аудита качества учебного процесса [6, 26].

Поскольку оценочные шкалы критериально-ориентированных тестов, как правило, или двухбалльные, или четырехбалльные (пятибалльные), дифференцирующая способность таких тестов относительно низкая, что позволяет использовать для перевода баллов за решенные ТЗ в итоговые оценки относительно простые априорные критерии, формулируемые и регулируемые разработчиком теста или группой экспертов на основе практического опыта и детерминированных математических моделей.

Если провести аналогию с теорией и техникой измерений, можно отождествить применение нормативно-ориентированных тестов с измерениями физических величин методами непосредственной оценки прибором с нелинейной функцией преобразования. Такие измерения могут быть как прямыми, когда измеряемой величиной являются знания испытуемого в какой-либо предметной области, так и косвенными, когда с помощью нормативно-ориентированных тестов опосредованно измеряются прирост знаний, обучаемость, темпы забывания изученного материала и другие параметры, характеризующие когнитивные способности или факторы когнитивной деятельности, т. е. реально латентные величины.

Применение критериально-ориентированных тестов может быть соотнесено с измерениями физических величин методами сравнения с мерой, где в качестве меры выступают критерии оценивания, численно выраженные в баллах, начисляемых за решение ТЗ, пороговыми значениями и диапазонами оценок. Ввиду низкой прецизионности меры (балльно-оценочной шкалы) использование критериально-ориентированных тестов для измерений каких-либо латентных величин нецелесообразно ввиду больших интервалов их оценок (больших относительных погрешностей) при расчетах результатов косвенных измерений. По той же причине статистическая обработка массивов эмпирических данных, как правило, ограничивается нахождением и сравнением средних значений, дисперсий, среднеквадратических отклонений отдельных выборок, попарных корреляционных моментов выборок, а при индивидуальном закреплении результатов – построением и анализом корреляционных диаграмм [28].

Поскольку критериально-ориентированное тестирование в большей степени направлено на проверку, чем на оценку, для интерпретации его результатов требуется не очень высокая дифференциация ТЗ по уровням трудности: от простых до определенного критериального уровня. Разработчики критериально-ориентированных тестов могут включать в них относительно малое число ТЗ повышенной трудности для проверки глубины освоения отдельных ДЕСО, которые также могут играть роль маркеров надежности теста [29].

И критериально-ориентированные, и нормативно-ориентированные тесты используются в системах мониторинга реализации образовательных программ: первые – как инструмент контроля, вторые – как инструмент исследования динамики процессов и факторов, влияющих на качество обучения.

Существуют различные классификации тестов, в разной степени поддерживаемые и развиваемые различными авторами.

По новизне методических принципов разработки и реализации тесты подразделяются на *традиционные* (гомогенные или гетерогенные), когда испытуемые отвечают на идентичные задания в одинаковое время и в одинаковых условиях и оцениваются по единой шкале и критериям, а также *нетрадиционные*: интегративные, адаптивные, многоступенчатые.

Гомогенный тест применяется для оценки учебных достижений или проверки остаточных знаний испытуемых по одной учебной дисциплине или ее части [30, 97]. *Гетерогенный тест* предназначен для оценки учебных достижений испытуемых по нескольким учебным дисциплинам, состоит из нескольких гомогенных субтестов и, как правило, не включает в себя ТЗ междисциплинарного содержания [6]. Такие тесты применяются для оценивания сформированности у обучающихся сложных компетенций, комплексной оценки уровня подготовки выпускников образовательной организации [31].

Интегративные тесты включают в себя ТЗ интегративного содержания, возрастающей трудности и, как правило, имеют целью итоговое оценивание учебных достижений: по окончании обучения по программам дополнительного профессионального образования или по программам высшего или специального профессионального образования, например накануне или в ходе государственной итоговой аттестации [22]. Правильные ответы на ТЗ интегративного теста требуют взаимосвязанных знаний по двум и более учебным дисциплинам. За счет интегративности и большей информационной емкости ТЗ их количество в интегративных тестах может быть существенно снижено по сравнению с гетерогенными.

В *адаптивном тесте* каждое последующее ТЗ или серия ТЗ (субтест) подбирается специальным адаптивным алгоритмом по результатам решения предыдущих заданий с учетом определенного заранее уровня их трудности. Уровень трудности первого субтеста выбирается испытуемым или задается адаптивным алгоритмом. В случае успешного решения предлагается более трудное ТЗ или субтест, и наоборот. Адаптивные алгоритмы обеспечивают индивидуализацию набора ТЗ, контролируют не только правильность, но и время их решения. У слабо подготовленных обучающихся формируется и поддерживается высокий уровень мотивации к выполнению теста за счет исключения трудных ТЗ. Адаптивные тесты, как и адаптивные автоматизированные алгоритмы измерения и контроля параметров техни-

ческих объектов, обладают потенциально минимальной погрешностью и потенциально высокой надежностью измерений [23, 24, 32].

При *пирамидальном адаптивном алгоритме* всем испытуемым сначала дается ТЗ средней трудности, затем в зависимости от ответа для каждого испытуемого уровень трудности ТЗ повышается или понижается. Еще в одном распространенном варианте реализации адаптивного тестирования испытуемые последовательно решают ТЗ возрастающей трудности, а затем алгоритм определяет трудность ТЗ оставшейся части теста индивидуально для каждого испытуемого [31, 33].

Многоступенчатые тесты состоят из нескольких субтестов возрастающей трудности, которые испытуемый должен проходить последовательно. Тестирование прекращается, если испытуемый не преодолел установленный порог минимального балла на очередном уровне.

Разные авторы [25–28, 33] исходя из целеполагания разработки тестов дополнительно классифицируют их следующим образом:

1) *диагностические* тесты, подразделяющиеся на тесты общих и специальных способностей, тесты обученности и учебных достижений;

2) *дидактические* тесты с особым отбором содержания ТЗ, не просто выявляющие знание или незнание учебного материала, но и позволяющие оценить полноту и системность знаний, их конкретность и общность, глубину и прочность, актуальность, гибкость, оперативность применения;

3) *по задачам тестирования*: тематические, итоговые, обучающие, развивающие, контролирующие остаточные знания;

4) *по средствам предъявления*: тесты на бумажных носителях с заполнением специальных бланков ответов; компьютерные тесты с фиксацией результата программными методами.

Реализация ТОЗУО компьютерными технологиями накладывает на структуру и содержание ТЗ, процедуры оценивания ряд специфических ограничений. Так, в компьютерных тестах ввиду ограниченных возможностей проверочных алгоритмов или лимита времени тестирования практически не используются ТЗ с большой трудоемкостью выполнения, ТЗ с множественными верными решениями, или решаемые в общем виде (без числового ответа, кроме ТЗ с выбором символьного выражения из предлагаемых вариантов), или творческого, поискового характера, ТЗ открытого типа, подразумевающие самостоятельную формулировку ответа испытуемым, либо исключается употребление синонимов и смысловых эквивалентов в ответах на прямые вопросы. Таким образом, компьютерное тестирование позво-

ляет достоверно оценивать знания и умения, характеризующие нижние уровни освоения обучающимися компетенций, или теоретическое содержание обучения и базовые практические умения по какой-либо дисциплине без претензий на достоверное оценивание более сложных результатов обучения.

В силу данных особенностей ТОЗУО на базе компьютерных технологий определим как КТОЗУО.

Прогресс в разработке компьютерных тестирующих программ существенно расширил возможности и сферы применения дидактических тестов в диагностике, мониторинге и аудите качества освоения образовательных программ, управлении качеством образования, повысил их конкурентоспособность среди инструментов КУДО [35–38].

Важной характеристикой теста, обеспечивающей дифференцируемость оценок, достоверность оценивания учебных достижений обучающихся, корректность анализа результатов и их сопоставления, является *время решения теста*. Оптимальное или предельное (максимально допустимое) время тестирования определяется эмпирически и (или) экспертными оценками и указывается для каждого теста. Обычно время тестирования полагается равным:

$$T = \overline{\Delta T}_L \cdot n_L + \overline{\Delta T}_C \cdot n_C + \overline{\Delta T}_T \cdot n_T, \quad (1.1)$$

где $\overline{\Delta T}_L$, $\overline{\Delta T}_C$, $\overline{\Delta T}_T$ – соответственно средние времена решения легких, средней трудности и трудных ТЗ, определяемые эмпирическим или экспертным способом; n_L , n_C , n_T – соответственно количества легких, средней трудности и трудных ТЗ.

Лимит времени может устанавливаться как для теста в целом, так и для каждого ТЗ, особенно легких, предполагающих быстрый выбор верного варианта ответа или быстрый устный счет, для уменьшения вероятности угадывания правильного ответа.

К важным характеристикам теста относится и *вариация тестовых баллов*, т. е. различие тестовых баллов у разных испытуемых, которую можно рассматривать как необходимое условие получения достоверных оценок учебных достижений обучающихся. Отсутствие вариации или ее относительно малые значения свидетельствует либо об одинаковом уровне знаний и умений тестируемых, либо о низкой *дифференцирующей способности* теста и его несостоятельности в дифференцировании знаний и умений испытуемых. Грубой оценкой вариации является интервал тестовых баллов – разность максимального и минимального значения баллов, набранных всеми испытуемыми, или приведенный интервал – отношение интервала к среднему значению набранных баллов. На величину интервала могут оказывать

существенное влияние аномальные статистические выбросы, прежде всего в отношении минимального балла, поэтому более корректной и удобной для расчетов мерой вариации результатов тестирования является дисперсия тестовых баллов. Ее применение для анализа и интерпретации результатов, в том числе в алгоритмах ранжирования уровней учебных достижений испытуемых, наиболее уместно для нормативно-ориентированного и адаптивного тестирования.

Дифференцирующая способность теста обеспечивается его формированием из ТЗ различной трудности.

Уровень трудности отдельных ТЗ и их представительство в тесте определяется целями тестирования. *Трудность теста* определяется суммарной трудностью всех его ТЗ.

Исторически простейшей мерой трудности некоторого j -го задания в дидактической квалиметрии была доля правильных ответов p_j на него. В некоторых моделях тестов наряду с p_j в качестве показателя трудности j -го задания применяется доля неправильных ответов на него: $q_j = 1 - p_j$.

Современная методология конструирования тестов в качестве меры *уровня трудности заданий* оперирует *логитом трудности задания*, определяемым натуральным логарифмом отношения доли данных испытуемыми неправильных ответов на ТЗ к доле правильных ответов на это же задание: $\ln \frac{q_j}{p_j}$. *Логит уровня подготовленности i -го*

учащегося определяется как $\ln \frac{P_i}{Q_i}$, где P_i и Q_i – соответственно доля данных им правильных и неправильных ответов [8, 39]. Сопоставление логарифмических оценок уровня знаний каждого испытуемого с уровнем трудности каждого ТЗ посредством их вычитания позволяет создавать программные средства индивидуализации обучения и контроля, в том числе адаптивного.

Тесты обладают *тестологическими характеристиками*, т. е. признаваемыми профессиональным сообществом измерительными качествами. Тестологические характеристики с высокой степенью достоверности могут быть оценены по итогам апробации теста на экспериментальной группе тестируемых или по мере накопления достаточно-го объема выборки в процессе практического применения теста. Постэмпирическое оценивание тестологических характеристик в первом случае требует сравнительно трудоемких исследований при условии обеспечения объема эмпирического материала и репрезентативности

выборки испытуемых, достаточных для корректного статистического анализа, а во втором случае сопряжено с частичной утратой актуальности результатов и сравнительно большим временем получения экспериментальных данных, пригодных для анализа, особенно с учетом возможной корректировки теста в целях повышения или поддержания его качества в процессе применения.

Эмпирические методы оценивания тестологических характеристик делают создание теста с заданными параметрами качества оценивания учебных достижений сложным, длительным и трудоемким процессом, включающим выбор математической модели конструирования теста, определяющей его структуру по заданным характеристикам качества, наполнение тестовой структуры ТЗ соответствующего содержания и трудности, проведение апробационного тестирования, измерение искомых параметров характеристик теста и их сопоставление с заданными, коррекцию теста и его дальнейшее совершенствование методом последовательных итераций до достижения соответствия заданным тестологическим свойствам.

Методологическими проблемами реализации данного подхода являются определение выборки тестируемых в процессе апробации с обеспечением репрезентативности по численности и уровням учебных достижений, неопределенность количественного описания характеристик теста, способов их измерения по итогам апробаций и критериев сравнения задаваемых и получаемых эмпирическим путем параметров, критериев сходимости итерационных процедур конструирования теста.

Тем не менее апробирование теста и эмпирическое оценивание его характеристик на предмет его возможной корректировки, несомненно, способствует повышению достоверности последующих результатов тестирования.

В практике разработки тестов возрастает роль априорных экспертных оценок их тестологических характеристик, а также оценок на основе аналогий и опыта применения тестов самими разработчиками. Достоверность таких оценок обеспечивают и математические модели оценивания характеристик теста, например надежности теста и его отдельных структурных компонентов (заданий, процедур формирования баллов и их отображения на шкалы оценок), а также имитационные модели тестируемых и воспроизведения тестовых алгоритмов [21, 40–43, 87].

При разработке ТОЗУО должны оптимально сочетаться априорные и апостериорные методы оценивания его качества. Первые, реа-

лизированные в математических моделях и экспертных оценках, существенно снижают трудоемкость и время разработки теста. Вторые позволяют окончательно и предметно судить о достижении поставленных разработчиками целей.

Наиболее значимыми характеристиками качества теста являются *надежность* и *валидность*. Данные характеристики взаимосвязаны и, как показывает анализ публикаций, зачастую наделяются теоретиками-тестологами и разработчиками тестов дублирующими признаками.

Идея достижения требуемой точности педагогических измерений заданиями теста теоретически базируется именно на надежности. Это вытекает из известного постулата о неизбежности погрешности любых измерений: измеряемая величина X не равна истинному значению T [31]. Под надежностью теста чаще всего понимается устойчивость результатов тестирования, т. е. способность давать одни и те же результаты при его применении к одинаковым выборкам испытуемых [20, 44].

Такая узкая трактовка надежности и ее сведение к единственному свойству противоречит теории надежности, в которой надежность является комплексной характеристикой, определяемой целым рядом рассчитываемых и экспериментально измеряемых параметров. Более подробно данное противоречие рассмотрено и разрешено в гл. 2 настоящей монографии.

Сведение надежности к повторяемости результатов тестирования определяет и упрощенные подходы к ее количественному описанию. Так, в качестве меры повторяемости результатов тестирования используется коэффициент корреляции между результатами двукратного тестирования одного и того же контингента испытуемых по эквивалентным вариантам тестов. О надежности (в смысле повторяемости оценок) тестов судят по степени сохранения ранговых позиций испытуемых. Иногда для определения надежности гомогенных тестов по коэффициенту корреляции используется метод расщепления, описанный в работах А. Анастаси, С. Урбина [45], Ю.М. Неймана, В.А. Хлебникова [39], М.Б. Челышковой [46]. Тест разделяют на две эквивалентные половины, затем вычисляют коэффициент корреляции r_t между результатами тестирования по двум половинам теста. При этом получается значение коэффициента корреляции только для половины теста, для целого теста оно определяется из соотношения

$$r'_t = \frac{2r_t}{1 + r_t}.$$

Валидность характеризует способность теста оценивать требуемые показатели. Создание теста может преследовать различные цели, поэтому, руководствуясь различными критериями целевой адекватности теста, валидность оценивают с разных позиций. Вследствие этого понятие валидности носит эклектичный характер, зачастую в него вкладываются, помимо целеполагающих свойств, признаки надежности, дифференцирующей способности, меры трудности теста. Так, валидность зависит от качества заданий теста, их числа, степени полноты и глубины охвата в них содержания обучения; от баланса и распределения заданий по трудности; от метода отбора заданий из общего банка, интерпретации результатов тестирования; от организации сбора данных, отбора выборки испытуемых [8, 47]. В [8] валидность ставится в зависимость от процедуры тестирования, включая выбор оптимального времени тестирования и попытки испытуемых угадать правильный ответ. Если первое действительно является регулируемой характеристикой теста, то второе может оказывать влияние на валидность (корректность, достоверность) результата тестирования, но не теста. При этом неудачные формулировки ТЗ, содержащие подсказки или некорректный набор дистракторов, способствуют угадыванию верных ответов и влияют на достоверность итоговых оценок.

Валидность оценивается на основе разнообразной информации и подразделяется на несколько типов:

диагностическую, отражающую способность теста дифференцировать испытуемых по изучаемому признаку, в частности относительно некоторой меры данного признака, например по результатам тестирования структурировать знания и умения испытуемых (в том числе относительно некой оптимальной модели результатов обучения);

прогностическую, определяющую «степень обоснованности и статистической надежности исследования измеряемого качества в будущем; возможность отбора учащихся по определенным признакам, например абитуриентов, способных успешно обучаться в вузе» [44];

эмпирическую, определяемую экспертными оценками специалистов (педагогами, учеными-экспертами, сотрудниками центра тестирования и др.);

конструктивную, свидетельствующую о теоретической обоснованности методики и соответствии результатов тестирования теоретическим прогнозам;

содержательную, определяемую адекватностью ТЗ проверяемому содержанию обучения;

функциональную, характеризующую соответствие ТЗ уровню усвоения контролируемых знаний;

критериальную, отражающую направленность теста на измерение соответствия учебных достижений заранее определенным критериям или эталонам, например образовательному стандарту (в отличие от диагностической валидности, количественной мерой критериальной валидности служат коэффициенты корреляции между показателями теста и заданной критериальной мерой).

1.3. Виды тестовых заданий

Объективности оценок теста способствует включение в него заданий разных видов.

Ряд специалистов наделяет ТЗ такими проверяемыми эмпирическим путем характеристиками, как трудность, дифференцирующая способность, вариативность, локальная независимость, технологичность и эффективность. Без такой апробации задания не могут быть тестовыми и являются *претестовыми*. Требование известной трудности является важнейшим системообразующим признаком ТЗ, на который ориентируются при его разработке. Опираясь категориями «легко» или «трудно», мерой которых является академический, педагогический и методический опыт, разработчик теста не только создает ТЗ примерной трудности, но и впоследствии трудность ТЗ корректирует, меняя его формулировку или способ задания условия, вид, набор дистракторов и др., доводя до необходимой ему категории трудности.

На наш взгляд, такие параметры, как технологичность, вариативность и локальная независимость ТЗ, могут быть определены априори, умозрительно – самим разработчиком теста или экспертами. Отчасти это относится и к дифференцирующей способности, а эффективность теста или ТЗ может быть оценена только по результатам завершенной апробации на достаточно большом количестве испытуемых, обеспечивающем состоятельность оценок данных характеристик. Часто имеющая место несостоятельность оценок эффективности, трудности и дифференцирующей способности ТЗ и ТОЗУО в целом, основанная на пробных тестированиях, подпитывает убежденность сторонников формирования тестов на основе двухпараметрической модели Раша [33, 48–50] в том, что несоответствие модели эмпирическим данным является проблемой практики, а не теории [51].

Виды ТЗ подразделяются на две группы: *задания в открытой форме (открытого типа)* и *задания в закрытой форме (закрытого типа)*.

В ТЗ открытого типа набор готовых ответов для выбора отсутствует, испытуемым необходимо самостоятельно вводить запрашиваемую информацию. Задания открытого типа подразделяются на *задания с дополнением* и *задания свободного изложения*. В первом случае, как правило, необходимо ввести от одного до трех слов или символов. В результате задание должно превратиться в истинное логическое высказывание или формулу. Проверка правильности ответа проводится путем сравнения вводимых слов или символов с эталоном верного решения, вариативность верного ответа исключена. При свободном изложении объем вводимой информации может быть значительно больше, на ответы не накладываются какие-либо ограничения, за исключением в отдельных случаях размеров поля ввода текста. Автоматизированная проверка ТЗ свободного изложения крайне затруднена, поэтому в КТОЗУО они используются в основном для проверки знаний дословных формулировок (например, пословиц, цитат из нормативных правовых актов, художественных произведений).

На практике чаще используются ТЗ закрытого типа, в которых испытуемому необходимо выбрать один или несколько ответов из предложенных вариантов или каким-либо образом упорядочить лингвистические или символичные конструкции. Варианты неверных ответов в тестовых заданиях с выбором одного или нескольких правильных ответов называются *дистракторами*. Вероятность случайного угадывания правильного ответа находится в обратной зависимости от количества дистракторов. Вместе с тем чрезмерное количество дистракторов приводит к увеличению времени решения задания и повышает вероятность случайной ошибки при выборе правильного ответа. Как правило, в ТЗВО количество дистракторов составляет от трех до пяти. Их меньшее количество может быть обусловлено спецификой задания, как, например, в тестах на знание Правил дорожного движения. Дистракторы должны отвечать принципам правдоподобия, равной привлекательности, стилистической однородности.

Ряд специалистов полагает, что формулировки дистракторов должны обеспечивать их выбор с одинаковой вероятностью (частотой). По нашему мнению, анализ неправильных ответов испытуемого представляет для педагога не меньший интерес, чем анализ соотношения верных и неверных решений ТЗ. Выбор заведомо неверного варианта ответа означает полное незнание содержания тестируемой

ДЕСО и может косвенно свидетельствовать о том, что часть успешно решенных вопросов теста была известна данному испытуемому заранее, т. е. стать основанием для перепроверки знания испытуемым контролируемого материала повторным тестированием или собеседованием. Неоднократный выбор испытуемым внешне состоятельных, но по сути абсурдных дистракторов, особенно из разных разделов содержания учебной дисциплины, уже однозначно говорит о несамостоятельности его верных решений заданий смежного содержания. Выбор испытуемым такого рода дистракторов должен оцениваться по-другому, нежели выбор вариантов ответа, наиболее близких к правильному, например частично верного или отличающегося от верного смысловыми нюансами. Учет данного обстоятельства в алгоритмах начисления баллов, по нашему мнению, поможет повысить не только объективность оценивания учебных достижений, но и надежность и дифференцирующую способность теста и его отдельных ТЗ.

Вероятность выбора дистракторов находится в обратной зависимости от уровня подготовленности испытуемых. Вместе с тем уменьшение частоты выбора дистракторов будет приводить к снижению дифференцирующей способности теста, а если такое уменьшение характеризуется монотонной временной зависимостью, то это свидетельствует о деградации теста или отдельных ТЗ и необходимости модернизации базы заданий. С целью повышения дифференцирующей способности ТЗ дистрактор, который никто не выбирает в качестве правильного ответа, необходимо откорректировать или заменить. Считается, что творчество разработчика теста состоит именно в создании неправильных, но очень правдоподобных ответов.

В качестве методического приема создания дистракторов М.Б. Челышкова рекомендует предъявлять обучающимся задания в открытой форме. Последующий анализ ошибок в представленных обучающимися ответах позволяет создать правдоподобные, с точки зрения испытуемых, дистракторы [6].

В.С. Аванесов определил принципы формулирования (подбора) ответов к заданиям с выбором единственного верного ответа [52, 53].

1. *Принцип противоречивости* (ответы подбираются взаимоисключающими, с использованием отрицания «не»).

Пример 1.1. Победа в Куликовской битве:

- а) привела к освобождению Руси от ига Золотой орды;
- б) не привела к освобождению Руси от ига Золотой орды.

2. *Принцип противоположности* (ответы формулируются как антонимы, допускаются переходные варианты от правильного ответа к взаимоисключающему).

Пример 1.2. После Куликовской битвы роль Московского княжества в отношениях Руси с Золотой ордой:

- а) возросла;
- б) уменьшилась.

3. *Принцип однородности* (варианты ответа относятся к одному множеству, смысловому или гомологическому ряду).

Пример 1.3. Перелом в Куликовской битве наступил в результате действий:

- а) передового полка;
- б) полка правой руки;
- в) полка левой руки;
- г) засадного полка.

4. *Принцип кумуляции* (включение предыдущих ответов в последующие с помощью союза «и» и запятых).

Пример 1.4. На стороне Мамае в Куликовской битве сражались воины:

- а) Золотой орды;
- б) Золотой орды и Рязанского княжества;
- в) Золотой орды, Рязанского княжества и Княжества Литовского.

5. *Принцип сочетания* (использование в вариантах ответа сочетаний одинаковых или близких по смыслу слов, знаков, терминов и др.).

Пример 1.5. На реке Калке в 1380 г. сражались войска:

- а) Мамае и Тамерлана;
- б) Тохтамыша и Дмитрия Донского;
- в) Мамае и Тохтамыша.

6. *Принцип градуирования* (использование градаций какой-либо характеристики для заданий с тремя и большим числом вариантов ответа).

Пример 1.6. После Куликовской битвы зависимость Московского княжества от Золотой орды:

- а) ослабла;
- б) не изменилась;
- в) усилилась.

7. *Принцип удвоенного противопоставления* (использование отрицания двух параметров для заданий с четырьмя ответами).

Пример 1.7. Победа в Куликовской битве способствовала:

- а) усилению Московского княжества и ускорению объединения русских земель;
- б) усилению Московского княжества и замедлению объединения русских земель;
- в) ослаблению Московского княжества и ускорению объединения русских земель;
- г) ослаблению Московского княжества и замедлению объединения русских земель [30].

Содержательная основа ТЗ разрабатывается на основе принципов *фасетности* и *импликации*.

Фасетность подразумевает создание конструкций (фасетов), состоящих из набора однородных элементов и используемых для формирования различных (параллельных) вариантов содержательной основы задания.

Пример 1.8. Поражение (победа) объединенных войск русских княжеств в битве на реке Калке (на Куликовом поле, на реке Сити) в 1223 г. (1238 г., 1380 г.) стало прологом 250-летнего ига Золотой орды (важным событием Западного (кипчакского) похода монголов, ускорило процесс объединения русских княжеств вокруг Москвы) [30].

В данном утверждении три фасета, из которых могут формироваться различные варианты ТЗ, а их количество – достигать 18. Фасетная структура ТЗ облегчает их разработку, экспертизу, корректировку, проверку выполнения, восприятие испытуемыми, позволяет уменьшить как количество случайных ошибок при ответе на ТЗ, так и вероятность искажения результатов тестирования за счет использования испытуемыми неразрешенных источников информации, упрощает сопоставление результатов испытуемых и повышает его объективность.

Импликация используется для проверки знаний причинно-следственных отношений и предполагает использование в явном или неявном виде логической связки «если... то...».

Пример 1.9. Вследствие увеличения объема денежной массы при неизменном объеме валового внутреннего продукта:

- а) увеличивается государственный долг;
- б) снижается покупательная способность национальной валюты;
- в) растет процентная ставка по кредитам [30].

Виды заданий закрытого типа:

- с альтернативными ответами;
- с выбором единственного верного ответа;
- с множественным выбором верных ответов;

на установление соответствий;
на установление последовательности;
с градуированными ответами;
с упорядочиванием дистракторов.

Тестовые задания с альтернативными ответами являются самыми простыми тестовыми заданиями закрытого типа. Вопросы, как правило, формулируются в форме согласия или несогласия с утверждением.

Пример 1.10. При переключении зеленого сигнала светофора с постоянного на мигающий водитель:

- а) обязан остановиться;
- б) не обязан остановиться [26].

Задания такого рода в тестах используются редко, поскольку вероятность случайного угадывания правильного ответа испытуемыми равна 0,5, т. е. предельно высока. Их рекомендуется использовать для ранжирования испытуемых по шкале «зачет – незачет», выявления или предварительного отсева наименее подготовленных испытуемых, проверки знания испытуемыми действий в ситуациях выбора альтернативы, как в приведенном выше примере. С целью уменьшения вероятности случайного угадывания для проверки знания, например одной ДЕСО, применяют серии заданий с альтернативами. ДЕСО считается освоенной, если на все задания в серии даны верные ответы. Однако для достижения той же цели более эффективны задания с большей вариативностью выбора ответов.

Структура ТЗВО является самой распространенной и хорошо известна испытуемым. Грамотно разработанные задания этого типа обеспечивают объективность оценивания учебных достижений при наличии эталона правильного ответа, не зависящего от субъективного мнения проверяющего. К достоинствам заданий с выбором единственно верного ответа следует отнести относительно низкую трудоемкость их разработки, простоту редактирования, малое время решения, универсальность применения, возможность проведения углубленного анализа освоения ДЕСО на основании статистики выбора испытуемыми дистракторов в качестве верных ответов, простоту и однозначность формирования оценочных шкал и алгоритмов начисления баллов.

Основными недостатками ТЗВО являются:

ограниченность диапазона тестируемых знаний и умений (как правило, ТЗВО не применяются для оценивания комплексных знаний и умений, умений творческого характера и т. п.);

относительно высокая вероятность угадывания испытуемым правильного ответа и провоцирование его к угадыванию самой формой отображения ТЗ на мониторе компьютера или бланке.

Для повышения мотивации испытуемых к отказу от попыток угадывания верного ответа можно, наряду с начислением баллов за верный ответ, отнимать за неверный, а за отказ от ответа штраф не применять [26].

В ТЗМВ испытуемый должен в представленном перечне вариантов ответа указать все верные.

Пример 1.11. Полностью на территории России расположены горные массивы:

- а) Уральские горы;
- б) Памир;
- в) Кавказ;
- г) Хибины;
- д) Алтай;
- е) Сихотэ-Алинь.

Тестовые задания с множественным выбором верных ответов занимают вторую позицию в частоте предпочтений разработчиков тестов [22, 30, 36, 54]. Они не уступают ТЗВО в тематическом охвате и простоте редактирования, и их большей распространенности препятствует неоднозначность трактовок разработчиками тестов и, как следствие, оценивания частично верных ответов (выбора, например, одного из двух верных ответов, одного верного и одного неверного и т. д.) [4, 9, 12]. ТЗМВ характеризуются существенно меньшей вероятностью угадывания полностью верного ответа и большей устойчивостью к деградации, чем ТЗВО, однако достоверность оценок КТОЗУО с их применением основывается на сложных критериях, которые должны учитывать как полностью, так и частично верные ответы. Такие критерии в значительной мере носят эвристический характер, базируются на субъективных оценочных аксиомах разработчика, поскольку, например, невозможно объективно сопоставить между собой 100% частично верных ответов и 50% полностью верных ответов одного и того же теста [6, 8, 54, 74]. ТЗМВ как объект удаленной подсказки не требует более высокой, по сравнению с ТЗВО, квалификации «третьих лиц».

В.С. Аванесов предлагает за полностью правильное решение начислять три балла, за каждую ошибку отнимать один балл, в случае допущения испытуемым более трех ошибок вычитать три балла [22]. Таким образом, в результате выполнения одного ТЗ количество воз-

можных баллов варьируется от нуля до трех при условии, что в задании три и более верных ответа. М.Б. Челышкова рекомендует давать один балл за полностью выполненное задание и ноль баллов за выбор хотя бы одного неверного варианта ответа [6]. В.Ю. Переверзев предлагает метод «частичного балла», в котором за каждый правильно выбранный ответ дается один балл, за неправильно выбранный ответ – ноль баллов. Штрафные баллы в этой методике не предусмотрены [54].

Разные способы начисления баллов влекут за собой различные методики перевода баллов в итоговые оценки, что усиливает влияние на достоверность результатов тестирования практического опыта разработчиков тестов, корректности применения ими математических моделей, эвристических и эмпирических подходов, тем самым повышая значимость этапа апробации тестов данного вида и принятия дополнительных мер по обеспечению репрезентативности экспериментальной выборки тестируемых. Таким образом, тесты на основе ТЗМВ более трудоемки и менее универсальны по сравнению тестами, составленными из ТЗВО, и предполагают более высокую квалификацию разработчиков ТОЗУО.

Задания на установление соответствий требуют от испытуемого найти соответствия между элементами двух множеств. Соответствие устанавливается на основании знаний (например, перевода иностранных слов, синонимов, антонимов), логических умозаключений или смысловых ассоциаций.

Структурно ТЗУС содержит инструкцию «установить соответствие» и два столбца. В левом столбце размещены элементы первого множества, в правом столбце – элементы второго множества. Количество элементов в правом столбце должно быть больше или равно количеству элементов в левом столбце. Количественное равенство элементов в обоих столбцах приводит к повышению вероятности угадывания верного решения, так как для последней пары элементов установление соответствия будет формальной процедурой.

Пример 1.12. Укажите столицы государств Южной Америки:

Эквадор	Монтевидео
Перу	Кито
Боливия	Каракас
Венесуэла	Лима
Уругвай	Сукре

Выполнение данного ТЗ может быть существенно затруднено добавлением в правый столбец еще двух названий столиц, например Сантьяго и Асунсьон.

Разновидностью ТЗУС являются задания с множественным соответствием, в которых каждому элементу левого столбца могут соответствовать несколько элементов правого столбца, а отдельным элементам левого столбца может не находиться соответствий.

Еще одна форма ТЗУС предполагает наличие третьего множества (столбца). В таких заданиях элементы первого множества сопоставляются с элементами второго и третьего множеств.

Для всех разновидностей ТЗУС, как и для ТЗМВ, характерна проблема неоднозначности трактовок верных решений и процедур начисления баллов за исчерпывающе и частично верные ответы.

Тестовое задание на установление соответствия охватывает существенно меньший круг дидактических единиц содержания различных предметных областей, так как для обеспечения приемлемого уровня вероятности угадывания полностью верного ответа количество парных соответствий должно быть более трех [1, 26]. Однако их применение в КТОЗУО в большей степени определяется трудностями синтаксиса заданий, опытом разработчиков, стремлением повысить дидактическую направленность ТЗ. В частности, установление соответствий широко практикуется в заданиях для дошкольников и младших школьников (в том числе в формате «картинка – слово», «арифметическое действие – ответ»), на уроках русского и иностранных языков («слово – перевод», «слово – синоним», «пропущенная орфограмма – буква» и т. д.), географии («государство – столица», «государство – государственный язык» и др.), химии («элемент – обозначение», «элемент – валентность» и др.). Простота оценивания таких ТЗ ограничивается установлением парных соответствий, при схеме соответствия более сложной, чем «один к одному», возникают проблемы с интерпретацией решений, как и при ТЗМВ [6, 8]. Само установление парных соответствий ведет к сокращению неизвестных испытуемому пар и повышает вероятность угадывания полностью верного решения ТЗ [26].

В ТЗУП испытуемому необходимо расположить в нужной последовательности элементы ответа. С помощью таких заданий удобно проверять знание и понимание испытуемыми небольших символьных выражений (формул), формулировок определений, понятий, терминов.

Пример 1.13. Для вычисления площади круга необходимо:

- 1) число π ;
- 2) разделить на;
- 3) возвести в квадрат;
- 4) умножить на;

- 5) четыре;
- 6) диаметр круга.

Ответ вводится как верная, по мнению испытуемого, последовательность номеров элементов ответа или, если позволяет интерфейс, путем перестановки таких элементов до образования читаемого предложения (формулы). Нередко, как в представленном примере, правильная комбинация цифр или элементов ответа не является единственной, что усложняет алгоритмы оценивания заданий. Не исключены и частично верные ответы, для которых так же, как и для ТЗМВ, возникает проблема неоднозначности оценивания.

Разновидностью ТЗУП является упорядочивание данных: расположение их в порядке возрастания или убывания количественных значений характеристики некоторого свойства группы объектов.

Пример 1.14. Расположите по последовательности правления российских императоров:

- 1) Александр I;
- 2) Николай I;
- 3) Павел I;
- 4) Петр III;
- 5) Екатерина II.

Пример 1.15. Расположите по прямой удаленности от Москвы города:

- а) Екатеринбург;
- б) Архангельск;
- в) Краснодар;
- г) Саратов;
- д) Симферополь.

Достоинством ТЗУП является однозначность оценивания решения, однако круг задач, охватываемый ТЗУП, сравнительно узок, а вероятность нахождения дословного (или близкого к дословному) правильного ответа в интернет-ресурсах велика [1, 38, 54]. Проблемой оценивания ТЗ на установление последовательностей исторических данных (дат, событий, персоналий), характеристик физических (расстояний, длин волн, частот, скоростей, плотностей веществ и т. п.), химических, биологических, географических объектов является равнозначность каждого элемента последовательности, что не позволяет сделать однозначный вывод о степени усвоения проверяемой ДЕСО при частично верном ответе [26, 30]. Вместе с тем поиск прямых ответов или эффективных подсказок в Интернете на подобные ТЗ

крайне затруднен, а широкие возможности формулирования заданий ограничивают возможности онлайн-подсказок от чатов ИИ.

Задания с градуированными ответами (с выбором наилучшего ответа, с градацией ответов) содержат ответы, которые имеют градацию по степени правильности. Задача разработчика заключается в нахождении и применении признака, позволяющего осуществить такую градацию. Максимальное количество баллов начисляется при совпадении последовательности ответов с эталонной последовательностью, установленной разработчиком задания. Случайное угадывание верных решений таких ТЗ маловероятно.

Пример 1.15. Относительная ширина полосы частот резонансного контура определяется значениями:

- 1) добротности контура;
- 2) емкости и индуктивности;
- 3) индуктивности и сопротивления;
- 4) емкости.

Конструирование ТЗ такого вида с четким критерием оценивания ответов является сложной задачей, требующей немалого искусства от составителя. Главная проблема, решаемая разработчиком теста, заключается не только в ограниченном количестве задач с множеством частично верных решений, но и в нечеткости критериев, определяющих градацию ответов, приводящих к субъективизму в оценивании решений. Для получения правильного решения испытуемый должен не просто знать проверяемый материал, но и воспроизводить при ответе логику составителя задания. Упрощение задания, например за счет уменьшения количества предлагаемых вариантов ответа до трех, существенно повышает вероятность угадывания их верной градации.

Введение в перечень предлагаемых испытуемому вариантов неверных ответов, наряду с уменьшенным количеством верных, с соответствующим изменением инструкции по выполнению ТЗ значительно упрощает составление задания и делает более прозрачными критерии его верного решения. Так, в примере 1.15 может быть добавлен пятый вариант ответа – «амплитуды резонансного напряжения». Такой перечень можно формировать с применением принципов кумуляции и сочетания, испытуемый должен градуировать ответы по степени полноты и достаточности перечня признаков и свойств объектов. В этом случае ТЗ сочетает в себе признаки ТЗГО и ТЗМВ.

Пример 1.16. Союзниками Германии в Первой мировой войне были:

- 1) Австро-Венгрия и Турция;
- 2) Австро-Венгрия, Турция, Болгария;
- 3) Австро-Венгрия, Турция, Япония;
- 4) Австро-Венгрия, Италия, Турция, Болгария.

При решении ТЗ испытуемый сначала исключает из рассмотрения неверные ответы, а затем выстраивает верные ответы в предполагаемой им последовательности. Задача оценивания ответа объединяет в себе сложность и неоднозначность оценивания ответов в ТЗМВ и ТЗГО.

Тестовые задания с градацией ответов, несмотря на их высокую устойчивость к динамике ошибок оценивания из-за угадывания верного ответа, относятся к числу редко используемых разработчиками КТОЗУО ввиду трудности подбора и формулирования заданий (далеко не для каждой ДЕСО можно придумать корректное ТЗГО), а также неоднозначности учета частично верных ответов (отсутствием единых критериев и алгоритмов начисления баллов за частично верные ответы) [8, 53, 54, 74]. Степень полноты и корректности предлагаемых на выбор вариантов ответа в различных ТЗГО даже одного КТОЗУО может качественно различаться, что затрудняет формализацию учета отклонений поучаемых градаций от эталонных: единый алгоритм начисления баллов и штрафов для отдельных ТЗ не обеспечивает достоверной (справедливой) оценки, а учет нюансов ТЗ в индивидуальных алгоритмах оценивания усложняет как использование доступных программных оболочек, так и разработку собственных программ, редактирование баз ТЗ и КТОЗУО в целом.

В [55, 56] приведена конструкция ТЗУД, по форме ответа тождественной ТЗУП или ТЗГО. Выполнение ТЗУД состоит из двух частей: 1) выбор единственного верного ответа из предложенных вариантов; 2) распределение дистракторов по степени близости к правильному ответу. Решение ТЗ дается в форме последовательности цифр: номер правильного ответа ставится на первую позицию, номер наиболее близкого к правильному – на вторую и т. д. ТЗУД выполняется как ТЗВО, в дополнение к выборочной проверке конкретного знания, возможно фрагментарного, и таким образом обеспечивает проверку понимания изученного материала в объеме как минимум ДЕСО. Тем самым ТЗУД обладает несопоставимо большими дидактическими возможностями по сравнению с наиболее распространенными конструкциями ТЗ [26, 30].

Пример 1.17. Расположите номера ответов в порядке близости ответов к верному. Закон Ома для участка цепи:

- 1) $Q = I^2 \cdot R \cdot t$;
- 2) $E = A/t$;
- 3) $I = U/R$;
- 4) $E = m/c^2$;
- 5) $I = E/(R + r)$.

Первую позицию вводимого числа должна занимать цифра 3. Ближайший к нему по достоверности – ответ под номером 5 (закон Ома для цепи с источником ЭДС). На третьей позиции стоит расположить ответ под номером 1 – одну из записей закона Джоуля-Ленца, объясняя свой выбор, во-первых, тем, что запись закона правильная, во-вторых, тем, что из оставшихся трех уравнений это единственное имеет отношение к электрическому току. Из оставшихся двух уравнений на четвертую позицию следует поставить то, что записано без ошибки. Решение ТЗ: 35142.

Пример 1.18. Расположите номера ответов в порядке близости ответов к верному. Кто из династии Рюриковичей был последним русским царем?

- 1) Дмитрий Донской;
- 2) Федор I;
- 3) Иван IV Грозный;
- 4) Гвидон Салтанович;
- 5) Симеон Бекбулатович.

Решение. В вопросе содержатся два систематизирующих признака: династия Рюриковичей и Русское царство, однако первый из них, судя по формулировке вопроса, является определяющим. Династию Рюриковичей в данном списке представляют Дмитрий Донской, Иван IV, Федор I, однако Дмитрий Донской был московским князем, но не великим князем всея Руси. Следовательно, первые три позиции последовательности: 231. Симеон Бекбулатович, не будучи Рюриковичем, был великим князем всея Руси. Гвидон Салтанович – персонаж «Сказки о Царе Салтане» А.С. Пушкина. Верный ответ: 23154.

По своей структуре и способу оформления решения ТЗУД похоже на ТЗГО, однако отличается от него принципами составления и решения. Основное отличие состоит в наличии дистракторов: кроме одного верного варианта ответа, все остальные являются неверными. Решение ТЗУД заключается в определении сначала верного ответа, что роднит этот вид ТЗ с ТЗВО, а затем в расположении дистракторов в порядке убывания от истины, что схоже с выполнением ТЗГО, толь-

ко по степени убывания истинности ответа необходимо располагать не частично верные ответы, а дистракторы. Конструирование ТЗ основано на подборе или формулировании разработчиком теста набора дистракторов, характеризующихся разной степенью принадлежности к соответствующим ДЕСО, предметной области, отрасли знаний.

Таким образом, структура данного ТЗ сочетает в себе достоинства ТЗГО и ТЗВО и существенно расширяет область задач с возможностью градации ответов.

Важным достоинством ТЗУД является его потенциальная устойчивость к угадыванию верных ответов, что определяет возможность многократного использования ТЗ данного вида в КТОУЗО без корректировки. В терминах теории надежности ТЗУД обеспечивает относительно большое по сравнению с ТЗ других видов время работы на отказ. Так, для задания с пятью вариантами ответа вероятность угадывания правильного ответа составляет $1/5! = 1/120$. С учетом возможностей компьютерной программы для создания тестов варьировать случайным образом очередность предлагаемых ответов, вводимая правильная комбинация не повторяется от одного сеанса тестирования к другому, что делает бесполезным запоминание правильного ответа для его сообщения другим испытуемым, а запоминание порядка следования и содержания всех вариантов ответов, приводимых в тесте в виде формул, графических объектов, предложений, и их передача от испытуемого к испытуемому по трудоемкости сопоставима с добросовестным изучением проверяемого содержания обучения.

Еще одним преимуществом ТЗУД, по сравнению с ТЗ других видов, является возможность дифференцировать оценки в зависимости от дидактических целей теста и глубины КУДО.

При изучении темы «Контроль знаний и умений с применением компьютерных тестов» в программе переподготовки «Преподаватель высшей школы» и в программе повышения квалификации профессорско-преподавательского состава «Инновационные образовательные технологии», реализуемых в Краснодарском университете МВД России, обучающимся в 2021 и 2025 гг. было предложено оценить по пятибалльной шкале некоторые качества различных видов ТЗ. Усредненные оценки приведены в табл. 1.1 (в числителях данные 2025 г., в знаменателях – 2021 г.). Здесь и далее в столбцах таблиц жирным шрифтом выделены медианы, подчеркиванием – максимальные значения, курсивом – минимальные значения.

Таблица 1.1

Оценки качеств тестов разработчиками в 2025 и 2021 годах

Вид ТЗ	Характеристики теста					Сумма баллов
	Досто- верность оценок	Устойчи- вость к де- градации	Обучаю- щий по- тенциал	Простота создания	Универ- сальность применения	
С альтернатив- ными ответами	2,8/2,7	2,6/2,8	3,0/3,0	<u>4,8/4,8</u>	4,3/ <u>4,5</u>	17,5/ 17,8
С единственным выбором	3,3/3,1	3,2/3,3	3,0/3,0	4,6/4,4	<u>4,6/4,3</u>	18,7/ 18,1
С множественным выбором	3,7/3,6	3,8/3,8	3,1/3,2	3,7/3,9	3,8/3,7	18,1/ 18,2
На установление соответствия	3,8/3,8	3,6/3,7	<u>4,1/3,6</u>	4,0/ 3,9	4,2/4,1	19,7/ 19,1
На установление последовательно- сти	3,9/3,9	3,5/3,6	3,9/3,6	3,3/3,4	3,5/3,5	18,1/ 18,0
С градуирован- ными ответами	4,3/4,4	4,3/4,2	4,0/ <u>4,0</u>	3,0/3,0	3,6/ 3,8	19,2/ 19,3
С упорядочива- нием дистракторов	<u>4,5/4,6</u>	<u>4,6/4,3</u>	3,9/3,9	3,3/3,1	3,6/ 3,8	19,9/ 19,7
Среднее значение для ТЗ всех видов	3,80/ 3,73	3,66/3,67	3,57/3,47	3,85/3,79	3,94/3,96	18,76/ 18,60

Тестовое задание с упорядочиванием дистрактов набрало максимальную сумму баллов по итогам обоих опросов за счет высокой, по мнению респондентов, потенциальной достоверности оценок знаний обучающихся, получаемых в тестах на его основе, устойчивости к деградации (обеспечение достоверности оценок при многократном применении теста) и обучающего потенциала (возможность для составителя теста формировать у обучающихся знания в процессе их тестирования). Ожидаемо одно из последних мест ТЗУД заняло по показателю простоты создания (трудоемкости). Обращает на себя внимание динамика ТЗУС: выход на второе место в условном рейтинге видов ТЗ, приближение вплотную к ТЗУД по сумме баллов за счет главным образом повышения респондентами оценки обучающего потенциала. Несмотря на разностороннюю критику ТЗВО, данный вид ТЗ заметно прибавил в суммарной оценке за счет высоко и прогрессивно оцененными респондентами простоты создания и универсальности применения. В целом по всем видам ТЗ возросли баллы их обучающего потенциала, простоты создания, достоверности оценивания,

что свидетельствует в том числе о большем совокупном практическом опыте конструирования КТОЗУО второй группы респондентов.

Респонденты также оценили по шкале от трех до пяти баллов (3 – не обязательно (не требуется); 4 – желательно; 5 – необходимо) степень востребованности различных качеств разработчиков тестов, характеризующих их квалификацию, при разработке ТЗ различных видов (табл. 1.2). В числителях приведены 2025 г., в числителе знаменателе – усредненные оценки опроса 2021 г.

Таблица 1.2

Оценки требований к квалификации разработчиков тестов в 2025 и 2021 годах

Вид ТЗ	Показатели квалификации						Сумма баллов
	Глубина знания предмета	Профессиональный кругозор	Опыт педагогической работы	Опыт создания тестов	Общий кругозор, эрудиция	Творческие способности	
ТЗАО	4,0/4,2	3,5/3,4	3,1/3,0	3,1/3,2	3,1/3,2	3,0/3,2	19,8/20,2
ТЗВО	4,4/4,6	3,6/3,8	3,7/3,8	3,6/3,2	3,5/3,3	3,5/3,4	22,3/22,1
ТЗМВ	4,6/4,7	4,0/4,1	4,0/4,2	3,9/3,7	3,6/3,5	3,6/3,4	23,7/23,6
ТЗУС	4,7/4,6	4,1/4,0	4,0/4,2	3,9/3,8	3,7/3,5	3,9/3,7	24,3/23,8
ТЗУП	4,6/4,8	3,8/3,8	4,1/4,2	4,3/4,1	3,6/3,5	4,1/3,9	24,5/24,3
ТЗГО	5,0/5,0	4,4/4,6	4,4/4,5	4,6/4,4	3,9/3,6	4,3/4,2	26,6/26,3
ТЗУД	5,0/5,0	4,6/4,7	4,5/4,5	4,6/4,5	4,2/4,3	4,5/4,4	27,4/27,4
Среднее значение для ТЗ всех видов	4,61/4,64	4,0/4,1	3,97/4,06	4,0/3,84	3,66/3,56	3,84/3,74	24,1/24,0

Как отмечено респондентами, разработка ТЗУД требует от преподавателей не только глубоких познаний в своем предмете (наряду с разработкой ТЗГО), но и творческих способностей и общего кругозора. Сравнительно высоким требованиям, по мнению преподавателей, должны соответствовать разработчики ТЗГО, ТЗУП и ТЗУС, причем для последних характерна наибольшая положительная динамика общих требований. Среди показателей квалификации наибольший прирост степени востребованности выявлен для опыта создания тестов, менее выражен – для общего кругозора и творческих способностей, опыт педагогической работы и профессиональный кругозор характе-

ризуются отрицательной динамикой оценок востребованности второй группы респондентов по сравнению с первой группой.

В табл. 1.3 приведены оценки востребованности различных видов ТЗ для КУДО по пятибалльной шкале (2 – исключено, 3 – допустимо, 4 – желательно, 5 – необходимо) тех же групп респондентов (в числителях – усредненные оценки опроса 2025 г., в знаменателях – усредненные оценки опроса 2021 г.).

Таблица 1.3

Оценки применимости различных видов тестовых заданий в 2025 и 2021 годах

Вид ТЗ	Применение в КУДО						Сумма баллов
	Текущий контроль		Промежуточная аттестация		Остаточные знания		
	самосто- ятельно	вместе с другими видами ТЗ	самосто- ятельно	вместе с другими видами ТЗ	самосто- ятельно	вместе с другими видами ТЗ	
ТЗАО	2,2/2,4	3,0/3,0	2,0/2,0	2,1/2,2	2,1/2,3	3,1/3,0	14,7/ 15,1
ТЗВО	4,1/3,9	4,6/4,5	3,4/3,5	4,7/4,6	4,0/4,2	4,8/4,7	25,6/ 25,4
ТЗМВ	3,4/3,7	4,3/4,0	3,2/3,3	4,2/4,1	3,4/3,7	4,3/4,2	22,8/ 23,0
ТЗУС	3,5/3,5	4,5/4,3	2,6/2,4	4,6/4,2	2,9/2,6	4,4/4,2	22,5/ 21,2
ТЗУП	2,7/2,6	3,9/3,7	2,2/2,3	4,0/3,9	2,3/2,5	3,8/4,0	18,9/ 19,0
ТЗГО	2,7/2,9	4,0/4,0	3,0/3,2	4,4/4,3	2,8/3,0	4,1/4,2	21,0/ 21,6
ТЗУД	3,1/3,2	4,4/4,1	3,8/4,0	4,6/4,5	3,1/3,1	4,4/4,3	23,4/ 23,2
Среднее значение для ТЗ всех видов	3,10/ 3,17	4,10/ 3,94	2,89/ 2,96	4,09/ 3,97	2,80/ 3,06	4,13/ 4,09	21,3/ 21,2

Из приведенных усредненных оценок следует, что, по мнению респондентов, наибольшей востребованностью и универсальностью применения в КУДО обладают ТЗВО, ТЗУД и ТЗМВ. Привлекательность применения всех без исключения видов ТЗ для разработчиков тестов возрастает при условии их использования совместно с другими видами ТЗ. Наибольшей самостоятельностью применения обладают ТЗВО (для контроля текущей успеваемости и остаточных знаний). КТОЗУО, составленный из ТЗУД, обладает хорошими перспективами применения в промежуточной аттестации. На фоне сравнительно стабильных суммарных баллов следует отметить существенный прирост

общей средней оценки ТЗУС и убыль среднего суммарного балла по оценкам второй группы респондентов для ТЗГО и ТЗАО.

В табл. 1.4 приведены обобщенные оценки потенциала применения ТЗ различных видов в КТОЗУО. Для расчета обобщенных показателей использованы данные из табл. 3 и суммарные средние баллы из правых столбцов табл. 1.1 и 1.2:

$$П_4 = П_1 - K_2 \cdot П_2 + K_3 \cdot П_3. \quad (1.2)$$

Здесь $П_1, П_2, П_3, П_4$ – показатель, соответствующий номеру таблицы в разделе (второе число в нумерации таблицы), K_2 и K_3 – нормировочные коэффициенты, обеспечивающие корректность сопоставления данных разных таблиц:

$$K_2 = \frac{5}{N_2}; \quad K_3 = \frac{5}{N_3},$$

где N_2 и N_3 – количество слагаемых (столбцов), из которых формируется используемый в формуле (1.2) показатель из таблиц 1.2 и 1.3 соответственно.

Таблица 1.4

Обобщенные оценки потенциала применения различных видов тестовых заданий в компьютерном тесте для оценивания знаний и умений обучающихся по опросам 2021 и 2025 годов

Вид ТЗ	Обобщенный показатель востребованности			
	В целом	Для текущего контроля	Для промежуточной аттестации	Для контроля остаточных знаний
ТЗАО	13,3/13,6	14,0/14,5	11,3/11,5	14,0/14,2
ТЗВО	<u>21,4/20,9</u>	<u>21,9/20,7</u>	<u>20,4/19,9</u>	<u>22,1/21,9</u>
ТЗМВ	17,3/17,7	17,6/17,8	16,8/17,0	17,6/18,3
ТЗУС	18,2/16,9	19,4/18,8	17,4/15,8	17,7/16,3
ТЗУП	13,5/13,6	14,2/13,5	13,2/13,3	13,0/14,0
ТЗГО	14,5/15,4	13,8/14,6	15,5/ 16,1	14,3/ 15,4
ТЗУД	16,6/16,2	15,9/15,1	18,1/18,1	15,9/15,4
Среднее значение для ТЗ всех видов	16,4/16,3	16,7/16,4	16,1/16,0	16,4/16,5

Безусловным лидером предпочтений преподавателей, имеющих различный опыт разработки КТОЗУО, является ТЗВО (максимальные баллы во всех столбцах). Наименьшими перспективами применения в КТОЗУО обладают ТЗ с альтернативными ответами, на установление последовательности и с градуированными ответами. Относительно высоко уровень востребованности для КУДО рецензенты оценили для ТЗ

с множественным выбором верных ответов (в особенности для контроля остаточных знаний), на установление соответствия (для текущего контроля) и с упорядочиванием дистракторов (для промежуточной аттестации). Следует отметить, что требования к квалификации разработчика ТЗ (см. табл. 1.2) учитываются выражении (1.2) со знаком «минус», что оказало заметное влияние на значения итогового показателя востребованности для ТЗГО, ТЗУД и ТЗУП. Явно выражен прирост оценок опроса 2025 г. по сравнению с опросом 2021 г. для ТЗВО, ТЗУС и ТЗУД, что обусловлено как растущим практическим опытом преподавателей в создании контрольно-измерительных материалов для текущего и итогового контроля, контроля остаточных знаний, так и объяснением принципов конструирования и демонстрацией возможностей ТЗ с усложненной структурой на методических семинарах и в процессе повышения квалификации профессорско-преподавательского состава. Причиной снижения обобщенных показателей востребованности для ТЗГО и ТЗМВ является одновременно сложность разработки ТЗ данных видов и неоднозначность оценивания частично верных ответов.

1.4. Реализация обучающего потенциала тестов и тестовых заданий

Нередко КТОЗУО, наряду с контролирующей функцией, реализуют и обучающую функцию. В некоторых контролирующе-обучающих программах завершение выполнения каждого ТЗ сопровождается демонстрацией правильного ответа, а иногда оценивается каждая попытка решения, т. е. испытуемому сообщается, правильно или неправильно он ответил. Такие подсказки снижают надежность КТОЗУО на основе ТЗ с усложненной структурой, так как количество I возможных исходов решения ТЗ, обратно пропорциональное вероятности угадывания верного ответа, в этом случае уменьшается. Так, например, для ТЗУС при угадывании ответа с «нулевым» знанием с подсказками:

$$I_0 = \sum_{i=2}^{n_c} i \leq n_c! . \quad (1.3)$$

Однако уже при $n_c = 3$ ТЗУС в алгоритмах тестирования с подсказками обеспечивает меньше вероятность ошибки второго рода, чем ТЗ с выбором одного верного варианта ответа из четырех, что обеспечивает КТОЗУО на основе ТЗУС дополнительный дидактический по-

тенциал. Условием получения удовлетворительной оценки при решении теста с подсказками может быть:

$$\frac{П}{П_{\text{зач}}} = \frac{П_{\text{в}} + П_{\text{н}}}{П_{\text{зач}}} \leq 1, \quad (1.4)$$

где $П$, $П_{\text{зач}}$, $П_{\text{в}}$, $П_{\text{н}}$ – соответственно общее количество совершенных попыток выбора правильного ответа, предельное количество попыток для получения зачета, количество верных попыток, количество неверных попыток.

Обучающей способностью, наряду с основной функцией оценивания знаний и умений, обладают и ТЗ в отдельности. Содержание ТЗ, их формулировки, наряду с подбором и формулировками дистракторов, элементов парных соответствий и других компонентов ТЗ, способны не только побудить познавательный интерес, направленный на поиск верного ответа, углубление знаний, расширение предметного кругозора, но и сформировать у испытуемого новое для него знание в процессе решения ТЗ. ТЗ могут содержать скрытые подсказки или наводящую на верный ответ информацию из смежной ДЕСО или группы таковых, которыми в принципе не сможет воспользоваться испытуемый, неуверенно ориентирующийся в содержании проверяемой дидактической единицы.

Наиболее простым и наглядным примером ТЗ, обладающего обучающей функцией, является ТЗУП. Верный ответ на ТЗ данного вида может быть получен, когда испытуемый, опираясь на собственные знания, составляет искомую последовательность частично, а затем, применяя логику, эрудицию, разрозненные фрагменты знаний, в том числе общих или из других предметных областей, выполняет задание полностью.

Вторым обстоятельством, обуславливающим сравнительно высокий дидактический потенциал ТЗУП, является возможность разработчика ориентировать ТЗ, в случае его успешного выполнения, на конкретизацию и детализацию знаний испытуемого – тем самым на их углубление и упрочнение. Так, например, искомая последовательность может представлять собой фазы или состояния физического объекта в различные моменты времени. Данная последовательность позволяет испытуемому исследовать количественно и качественно динамику процесса. Для реализации обучающих функций ТЗ элементы последовательности могут представлять собой этапы сложных химических реакций, эволюции организмов, исторических событий и процессов. Дополнительный эффект может быть достигнут при

использовании испытуемым изображений, графических построений при решении указанных ТЗ.

Еще одним фактором формирования дидактического потенциала ТЗУП является возможность адаптации уровня трудности одинаковых по содержанию ТЗ к уровню подготовки испытуемых и применения таких ТЗ в процедурах адаптивного тестирования [23, 24, 32, 33]. Например, ТЗ может быть сформулировано: «Расположите участки видимого диапазона по убыванию длины волны...» или «Расположите цвета радуги по порядку...». ТЗ во второй формулировке может как быть предназначено для проверки общих знаний, так и являться подготовительным к выполнению ТЗ в первой формулировке. И наоборот, если испытуемый затрудняется выполнить ТЗ в первой формулировке, ему в адаптивном тесте может быть предложена упрощенная формулировка, приводящая к качественно верному ответу.

В [96] нами рассмотрен такой аспект реализации дидактического потенциала ТОЗУО, как организация коллективной проектной деятельности обучающихся по разработке ТЗ и тестов для контроля учебных достижений. Среди принципов, на которых строятся коллективные учебные проекты, наибольшего внимания заслуживают [104–107]:

- совместная созидательная работа;
- получение реальных общественно полезных результатов;
- наглядность промежуточных и окончательных итогов работы;
- преемственность, перспектива продолжения (внедрения результатов) работы;
- сочетание индивидуальных и групповых форм работы, индивидуальной и коллективной ответственности за результат;
- обеспечение возможностей самореализации каждому исполнителю проекта.

Выбор нами разработки ТОЗУО в качестве учебного проекта обусловлен следующими соображениями:

1) корректное составление ТЗ, включая, собственно, саму формулировку вопроса или задания, подбор вариантов ответа, требует от составителя глубоких знаний проверяемой ДЕСО и как минимум элементарных знаний в целом содержания темы и ранее изученного содержания дисциплины (тем самым учебный проект оправдывает свое дидактическое предназначение и реализует функцию стимулирования самостоятельной внеаудиторной работы обучающихся в выбранной предметной области);

2) для конструирования тематических и дисциплинарных тестов необходим правильный выбор контролируемых ДЕСО и понимание их

взаимных приоритетов, что подразумевает понимание результатов обучения и их иерархии, логики построения контролируемого содержания, системы взаимосвязи ключевых понятий и фактов (таким образом учебный проект призван обеспечить формирование у его исполнителей системных знаний о структуре контролируемого содержания и упрочить запоминание его основ и тезауруса);

3) составление ТОЗУО и отдельных ТЗ требует от составителя знаний в области тестологии, в особенности прикладной педагогической квалитметрии, ее современных достижений, тенденций и перспектив, что обеспечивает формирование в ходе выполнения учебного проекта комплекса дополнительных знаний и умений, способных оказать положительное влияние на будущую учебную и профессиональную деятельность участников проекта;

4) разработка оценочных шкал тестов опирается как на качественный анализ результатов педагогических экспериментов, так и на статистическую обработку результатов тестирования, реализующую методы теории вероятностей и математической статистики, корреляционного анализа, статистического прогнозирования, что обеспечивает предметное формирование и закрепление знаний и умений в данных предметных областях;

5) процесс разработки ТОЗУО позволяет, в зависимости от состава и квалификации исполнителей, вариативно сочетать индивидуальные и групповые формы работы, консультирование и самостоятельную работу обучающихся над заданиями проекта;

6) составление надежных ТЗ, в особенности с усложненной структурой, требует от разработчиков не только предметной, но и общей эрудиции, знаний в смежных областях, что обеспечивает дополнительную реализацию развивающей функции учебного проекта;

7) поскольку ТОЗУО в подавляющем большинстве реализуются в специальных программных оболочках, работа над учебным проектом призвана углубить компетенции исполнителей в сфере информационных технологий;

8) очевидная практическая польза от реализации проекта заключается в формировании ФОС или его части по ряду учебных дисциплин с последующим применением для различных видов контроля и оценивания учебных достижений обучающихся в том числе в системах внутреннего мониторинга качества образовательного процесса.

По итогам реализации проекта по созданию обучающимися ТОЗУО по шести предметным областям (по четырем – в формате обучения и поэтапного усложнения ТЗ и расширения базы заданий,

по двум – в формате самостоятельной разработки исполнителями в малых группах) обучающиеся – исполнители проекта выделили следующие наиболее важные для себя результаты его выполнения (в порядке значимости, определяемой долей выбранных участниками проекта предлагаемых вариантов) [96]:

- получение новых знаний и умений по составлению ТЗ и ТОЗУО;
- развитие навыков подготовки докладов и их мультимедийного сопровождения;

- освоение новых программных продуктов;

- реализация возможности проявить свои знания и способности в коллективной работе;

- получение новых прикладных знаний в области теории вероятностей и математической статистики;

- повышение уровня знаний по предметным областям, по которым разрабатывались ТОЗУО;

- освоение методик самостоятельного опережающего изучения предмета.

1.5. Особенности разработки и применения тестов для адаптивного оценивания знаний

Несмотря на сложность реализации и трудоемкость разработки, адаптивные ТОЗУО востребованы в различных контрольно-оценочных мероприятиях, так как обеспечивают наивысшую объективность КУДО, входного контроля знаний для выяснения начальных условий обучения и других задач педагогической квалитметрии, решаемых с помощью тестов. Востребованность технологий адаптивного тестирования в современных условиях дополнительно обусловлена необходимостью обеспечения достоверности оценивания знаний в условиях несанкционированного применения испытуемыми новейших телекоммуникационных средств, интернет-технологий и ТИИ в ходе решения ими ТЗ. Расширенному применению адаптивных тестов препятствует отсутствие соответствующих алгоритмов, реализованных в коммерческих программных продуктах, с формированием из общей базы ТЗ тестов различного уровня трудности, учитывающего результаты предыдущих тестирований в рамках текущего сеанса. Параметрами, влияющими на трудность теста, могут быть не только трудность отдельных ТЗ, определенных экспертным или экспериментальным путем, но и объемы выборки ТЗ, значения диапазонов оце-

ночных шкал и время тестирования – время решения как теста в целом, так и отдельных ТЗ. Совокупное влияние таких параметров на трудность теста сложно поддается моделированию. Как следствие, это порождает проблемы с априорным определением трудности теста в процессе его разработки с учетом выбранной тактики адаптивного тестирования.

Адаптивные алгоритмы тестирования по способу оценивания знаний в каждом тесте условно можно разделить на две группы: с применением дифференцированных оценочных шкал (с дифференцированным оцениванием) и с установлением порогов положительной оценки (с недифференцированным оцениванием). Алгоритмы, относящиеся к первой группе, содержат, как правило, сравнительно мало (две-три) однотипных оценочных процедур (итераций), однако четырехбалльные оценочные шкалы в каждой процедуре требуют взаимного согласования. Количество последовательных оценочных процедур в алгоритмах второй группы достигает четырех, однако в данных процедурах критерий положительной недифференцированной оценки оперирует лишь единственным пороговым значением, что существенно упрощает их взаимное согласование, например на основе статистической модели Раша [48, 50].

Наиболее эффективные алгоритмы с дифференцированным оцениванием могут основываться на одном из двух обобщенных сценариев.

1. Первичное оценивание – сортировка испытуемых по категориям в зависимости от полученных оценок – окончательное оценивание с применением тестов, трудность которых соответствует категориям испытуемых, – дополнительное (при необходимости) тестирование для испытуемых с несогласованными оценками первичного и окончательного тестирования.

Несогласование проявляется в существенных (на 30% и более) различиях доли решенных ТЗ факультативного уровня трудности: легких ТЗ в тестах с преобладанием трудных и средней трудности ТЗ, предназначенных для подтверждения отличных и хороших оценок, или трудных ТЗ в тестах с преобладанием легких и средней трудности ТЗ, предназначенных для подтверждения первичных удовлетворительных оценок.

Итоговой является оценка окончательного или дополнительного (в случае его применения) тестирования.

2. Первичное оценивание – сортировка испытуемых по группам – повторное тестирование (уточнение первичной оценки) с уче-

том категории испытуемых – контрольное тестирование на основе средневзвешенных оценок первичного и повторного тестирования.

В данном алгоритме объем и трудность контрольного теста определяется мерой согласованности оценок первичного и повторного тестирования, характеризующейся величиной их отклонения от средневзвешенного значения. Поскольку трудность повторного теста адаптирована к выявленному первой оценочной процедурой уровню знаний испытуемых, при вычислении средневзвешенного значения оценке, полученной при повторной проверке, присваивается больший весовой коэффициент.

Примерные сценарии реализации алгоритмов с недифференцированным оцениванием можно разделить на три группы:

- с поэтапным повышением трудности;
- с поэтапным понижением трудности;
- с начальным тестом средней трудности (пирамидальный алгоритм [31, 33]).

Алгоритмы с поэтапным повышением трудности предполагают минимальную трудность первого теста, затем ее повышение с отсеиванием на каждом этапе испытуемых, не преодолевших порог перехода на следующий уровень трудности, и тем самым разделение испытуемых на четыре группы, соответствующие четырехбалльной шкале оценивания. Очевидно, для такого разделения достаточно трех уровней трудности тестов, т. е. количество уровней трудности на один меньше, чем количество предусмотренных шкалой оценок.

Более сложная реализация алгоритмов данного типа предполагает перепроверку достигнутого испытуемыми уровня трудности, например для теста с поэтапным повышением трудности (рис. 1.1):

I) оценивание тестом низкого уровня трудности $H1^*$, ориентированного в основном на проверку базовых элементов содержания, включающего в себя, наряду с ТЗВО, тестовые задания усложненной структуры – ТЗМВ, ТЗУС, ТЗУП;

IIa) повторное тестирование испытуемых, не достигших порога положительной оценки, тестом такого же низкого уровня содержательной трудности $H1$, включающим в себя минимизированное, по сравнению с предыдущим тестом, количество ТЗ усложненной структуры (испытуемые, не набравшие достаточного количества баллов для преодоления порогового значения получают оценку «неудовлетворительно», набравшие достаточное количество баллов – «удовлетворительно»);

IIб) оценивание испытуемых, решивших тест трудности $H1^*$, тестом среднего уровня трудности $C1^*$ с определенной долей ТЗ усложненной структуры, включая ТЗГО и ТЗУД;

IIIа) тестирование испытуемых, не решивших тест с уровнем трудности $C1^*$, тестом среднего уровня трудности $C1$ с минимизированной долей ТЗ усложненной структуры, не содержащим ТЗУД и ТЗГО (знания испытуемых, решивших тест с уровнем трудности $C1$, оцениваются на «хорошо»; не прошедшим испытание данным тестом предлагается тест с уровнем трудности $H1$ (этап IVа), и по результатам его решения выставляется оценка «удовлетворительно» или «неудовлетворительно»);

IIIб) тестирование испытуемых, решивших тест с уровнем трудности $C1^*$, тестом повышенной трудности (уровень $B1^*$), не менее двух третей ТЗ которого имеют усложненную структуру, включая ТЗУД и ТЗГО, и решение испытуемыми, не набравшими достаточного количества баллов для преодоления порогового значения теста $B1^*$, теста повышенной трудности (уровень $B1$) с минимизированным количеством ТЗУД и ТЗГО (этап IVб) (знания испытуемых, решивших данный тест, оцениваются на «отлично»);

Vа) решение испытуемыми, не прошедшими этап тестирования IVб, теста с уровнем трудности $C1$, по результатам решения которого испытуемым выставляется оценка «хорошо» или «удовлетворительно».

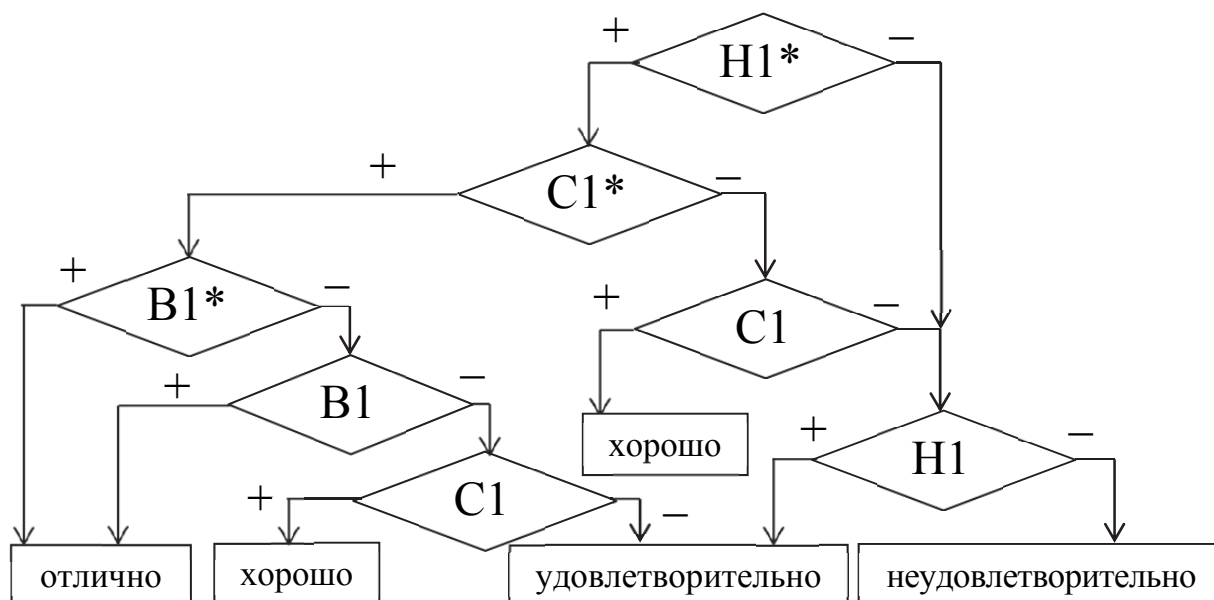


Рис. 1.1. Схема алгоритма адаптивного тестирования с поэтапным повышением трудности тестов и перепроверками

В данном алгоритме используется шесть отличающихся трудностью тестов, а количество этапов тестирования для испытуемых составляет от двух до пяти. Так, для получения оценки «неудовлетворительно» может быть затрачено от двух до четырех тестов, для получения оценки «удовлетворительно» – от двух до пяти, «хорошо» – от трех до пяти, «отлично» – от трех до четырех тестов. Таким образом, потенциально максимальную трудоемкость представляет оценивание знаний испытуемых на «хорошо» и «удовлетворительно», что объясняется необходимостью двустороннего отграничения данных оценок: «хорошо» от «отлично» и «удовлетворительно», «удовлетворительно» от «хорошо» и «неудовлетворительно».

Рассмотрим более простой алгоритм с начальным тестом низкого уровня трудности, оперирующий пятью тестами (рис. 1.2):

I) оценивание тестом с уровнем трудности $H1^*$;

IIa) повторное тестирование испытуемых, не выполнивших успешно первый тест, тестом с уровнем трудности $H1$ (выставление оценок «удовлетворительно» и «неудовлетворительно» испытуемым, соответственно решившим и не решившим второй тест);

IIб) оценивание испытуемых, решивших тест с уровнем трудности $H1^*$, тестом среднего уровня трудности $C1^*$;

IIIa) решение теста с уровнем трудности $H1$ испытуемыми, не решившими тест с уровнем трудности $C1^*$, с оцениванием их знаний по итогам решения теста «удовлетворительно» или «неудовлетворительно»;

IIIб) решение теста с высоким уровнем трудности $B1^*$ испытуемыми, успешно выполнившими тест с уровнем трудности $C1^*$ (выставление оценок «отлично» испытуемым, решившим данный тест);

IV) решение теста с уровнем трудности $C1$ испытуемыми, не решившими тест с уровнем трудности $B1^*$, и оценивание их знаний по итогам выполнения теста на «хорошо» (испытуемых, набравших количество баллов не менее порогового значения) или «удовлетворительно» (испытуемых, не набравших минимально необходимого количества баллов).

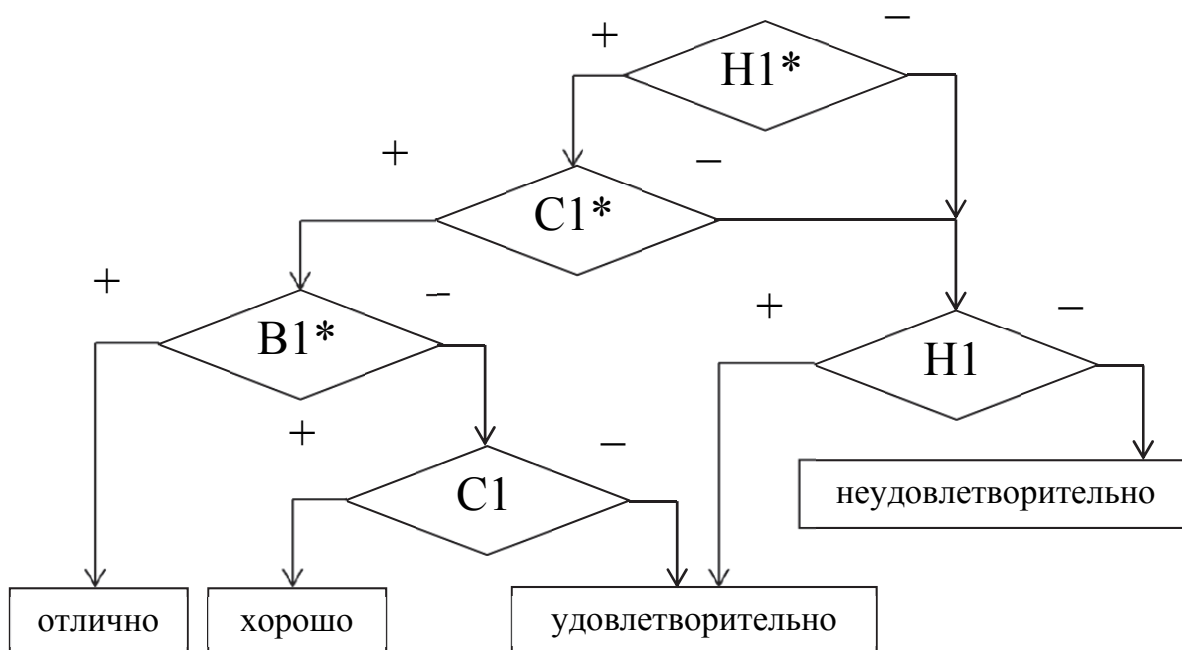


Рис. 1.2. Схема алгоритма адаптивного тестирования с поэтапным повышением трудности, предполагающего пять уровней трудности

Наиболее простой алгоритм с поэтапным повышением трудности и перепроверкой использует четыре отличающиеся по трудности теста (рис. 1.3):

I) оценивание тестом с уровнем трудности $H1^*$;

IIa) оценивание тестом с уровнем трудности $H1$ испытуемых, не выполнивших успешно первый тест, с выставлением оценок «удовлетворительно» или «неудовлетворительно» по итогам повторного тестирования;

IIб) оценивание испытуемых, решивших тест с уровнем трудности $H1^*$, тестом с уровнем трудности $C1^*$ (оценивание знаний испытуемых, не справившихся с данным тестом, на «удовлетворительно»);

III) решение теста с уровнем трудности $B1^*$ испытуемыми, решившими тест с уровнем трудности $C1^*$ (выставление оценок «отлично» или «хорошо» в зависимости от достижения испытуемыми порогового балла).

В данном алгоритме перепроверяется лишь оценка «неудовлетворительно», а оценки «хорошо» и «отлично» формируются с наивысшей трудоемкостью (три этапа тестирования).

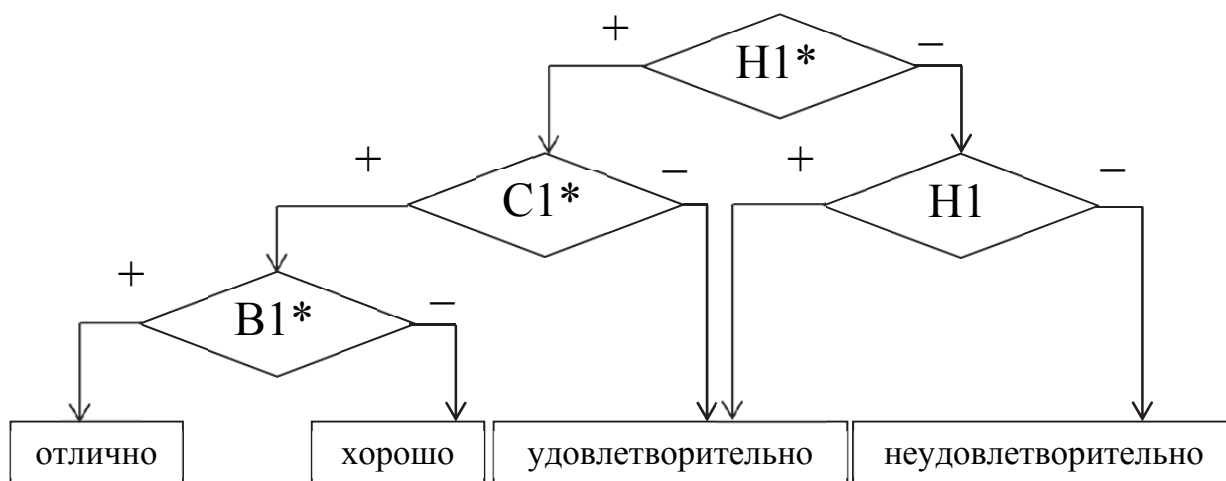


Рис. 1.3. Схема простейшего алгоритма адаптивного тестирования с поэтапным повышением трудности тестов и перепроверкой

Для адаптированного тестирования с поэтапным понижением трудности и перепроверкой возможен следующий алгоритм его реализации (рис. 1.4):

I) оценивание знаний испытуемых тестом с уровнем трудности B1*;

IIa) контрольное оценивание испытуемых, решивших тест с уровнем трудности B1*, тестом с уровнем трудности B1 (повторно преодолевшие установленное пороговое значение набранных баллов оцениваются «отлично»).

IIб) решение теста с уровнем трудности C1* испытуемыми, не решившими тест с уровнем трудности B1*;

IIIa) решение теста трудности C1 испытуемыми, решившими тест трудности C1* на этапе IIб, и испытуемыми, не решившими тест трудности B1 на этапе IIa, с оцениванием решивших и не решивших тест трудности C1 на «хорошо» и «удовлетворительно» соответственно;

IIIб) решение теста трудности H1 или H1* испытуемыми, не решившими тест трудности C1* на этапе IIб (оценивание испытуемых по результатам решения данного теста на «удовлетворительно» или «неудовлетворительно»).

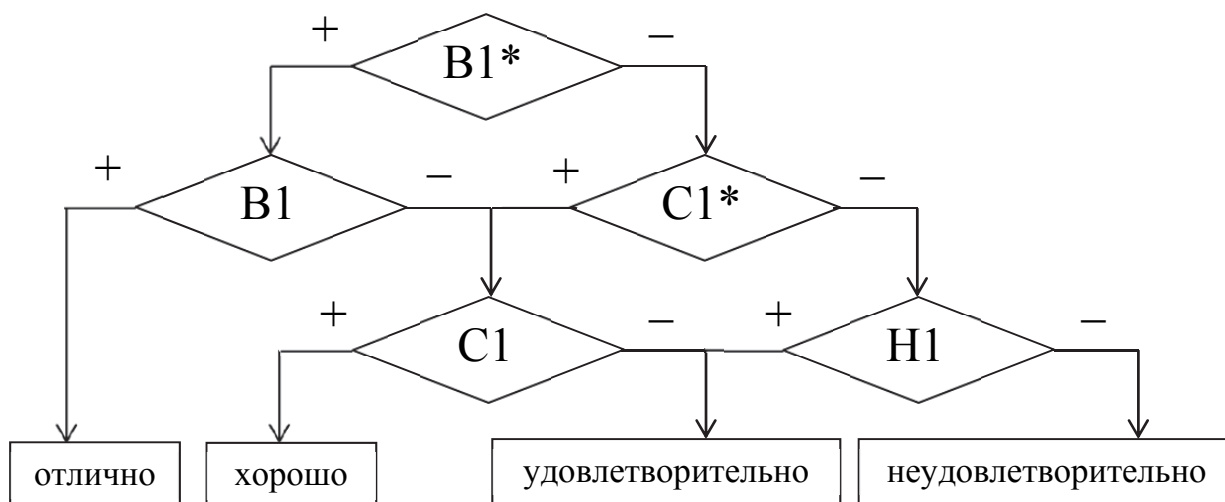


Рис. 1.4. Схема алгоритма адаптивного тестирования с поэтапным понижением трудности и перепроверками

Применение перепроверки результатов в алгоритме тестирования с начальным тестом среднего уровня трудности может быть представлено следующей последовательностью этапов (рис. 1.5):

I) оценивание испытуемых тестом с уровнем трудности $C1^*$;

IIa) решение испытуемыми, успешно прошедшими этап I, теста с уровнем трудности $B1^*$;

IIб) решение испытуемыми, не преодолевшими порог успешного прохождения этапа I, теста с уровнем трудности $H1^*$;

IIIa) контрольное тестирование испытуемых, успешно прошедших этап IIa, с применением теста трудностью $B1$ и оценивание решивших и не решивших данный тест на «отлично» и «хорошо» соответственно;

IIIб) решение испытуемыми, потерпевшими неудачу на этапе IIa, и испытуемыми, успешно прошедшими этап IIб, теста с уровнем трудности $C1$ (оценивание решивших и не решивших данный тест на «хорошо» и «удовлетворительно» соответственно);

IIIв) решение испытуемыми, потерпевшими неудачу на этапе IIIб, теста с уровнем трудности $H1$ с выставлением им оценок «удовлетворительно» или «неудовлетворительно».

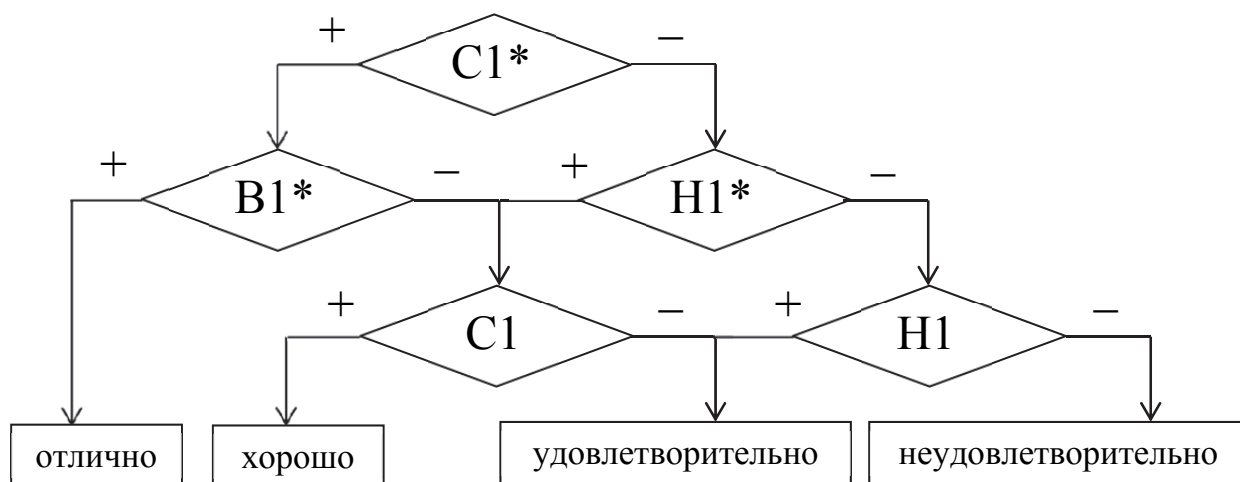


Рис. 1.5. Схема алгоритма адаптивного тестирования с начальным тестом среднего уровня трудности

Повышение уровня трудности КТОЗУО может осуществляться не только включением в их состав содержательно и структурно сложных ТЗ, но и повышением значения порога положительной оценки и сокращением лимита времени на решение теста или его отдельных заданий. Последний прием наиболее целесообразно применять для ТЗВО в тестах с относительно большим количеством ТЗ данного типа (уровни трудности Н1, С1), а также тестах, применяемых для подтверждения предварительных оценок или «реабилитации» испытуемых, потерпевших неудачи в решении тестов более высоких уровней трудности. Уменьшение лимита времени в данном случае призвано компенсировать наиболее очевидный недостаток ТЗВО в аспекте их надежности – подверженность угадыванию испытуемым верного ответа или получению его в процессе решения путем исключения испытуемым из рассмотрения отдельных дистракторов на основе собственных знаний, логических умозаключений, допускаемых составителями тестов подсказок, содержащихся в формулировках вопросов или возникающих при решении других ТЗ. Как следствие, испытуемый может дать верный ответ, изначально его не зная, что повышает реализуемость обучающего потенциала теста в ущерб выполнению задачи объективного контроля знаний. Уменьшение времени, отводимого на выполнение ТЗМВ и ТЗУС, по нашему мнению, должно сопровождаться смягчением критериев оценивания частично верных ответов, т. е. начислением баллов за неполные или частично верные ответы.

2. МАТЕМАТИЧЕСКИЕ МОДЕЛИ ОЦЕНКИ НАДЕЖНОСТИ КОМПЬЮТЕРНЫХ ТЕСТОВ ДЛЯ ПРОВЕРКИ ЗНАНИЙ И УМЕНИЙ ОБУЧАЮЩИХСЯ

2.1. Модели априорной оценки качества теста и достоверности результатов тестирования учебных достижений

Одна из первых проблем, с которой сталкиваются разработчики ТОЗУО, – определение оптимального (минимально достаточного) обеспечивающего объективность КУДО количества ТЗ, от которого напрямую зависит трудоемкость разработки теста. Данная проблема актуализировалась в связи с необходимостью формирования ФОС для проверки компетенций обучающихся и выпускников образовательных организаций высшего образования в соответствии с требованиями действующих федеральных государственных образовательных стандартов.

Корректность решения задачи определения количества ТЗ в тесте и в базе заданий в значительной степени зависит от корректности математической модели оценивания качества теста.

Различные подходы к оцениванию качества теста сходятся на том, что оно определяется достоверностью результатов тестирования, т. е. адекватностью соотнесения оценок, полученных по итогам тестов, с реальными знаниями и умениями испытуемых. Идеальной ситуацией является тождественность оценок тестов $O(T_T)$ в момент времени T_T неким безошибочным, абсолютно достоверным оценкам знаний и умений обучающихся $O^*(T_0)$, полученным независимыми способами в период времени оценивания:

$$T_0 \in T_T \pm \Delta T_{\text{и}}, \quad (2.1)$$

где $\Delta T_{\text{и}}$ – временной интервал, в течение которого возможно подлежащее учету (например, изменением времени решения ТЗ, корректировкой значения порогов и интервалов итоговых оценок или алгоритма подсчета баллов) изменение знаний и умений обучающихся (приращение вследствие их реконструкции и воспроизводства новых или убыль вследствие забывания) [38, 54, 58, 59].

Если предположить, что оценки $O^*(T_0)$ характеризуются абсолютной погрешностью:

$$\Delta O^*(T_0) = O^*(T_0) - O_{\text{и}}(T_0), \quad (2.2)$$

где $O_{\text{и}}(T_0)$ – некая идеальная оценка учебных достижений обучающихся, интерпретируемая, например, как математическое ожидание

частных оценок $O_i^*(T_o)$, полученных наиболее достоверным способом оценивания:

$$O_{\text{и}}(T_o) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N O_i^*(T_o), \quad (2.3)$$

тогда в широком смысле достоверность результатов тестирования может характеризоваться следующим условием:

$$O_{\text{т}}(T_o) \in [O_i^*(T_o) - \Delta O^*(T_o); O_i^*(T_o) + \Delta O^*(T_o)]. \quad (2.4)$$

В качестве дополнительного условия может рассматриваться неравенство

$$\Delta O^*(T_o) < \Delta O_{\text{т}}(T_o), \quad (2.5)$$

где

$$\Delta O_{\text{т}}(T_o) = O_{\text{т}}(T_o) - O_{\text{и}}(T_o). \quad (2.6)$$

С другой стороны, опираясь на теоретические достижения тестологии, обобщение опыта педагогов-практиков и собственные достижения в конструировании контрольно-измерительных материалов, разработчик тестов в своем стремлении повысить достоверность оценок тестирования способен достичь условия:

$$O_{\text{т}}(T_o) \pm \Delta O_{\text{т}}(T_o) \in [O_i^*(T_o) - \Delta O^*(T_o); O_i^*(T_o) + \Delta O^*(T_o)], \quad (2.7)$$

которое может служить критерием достоверности результатов тестирования в узком смысле.

Рассмотренные критерии носят апостериорный характер, т. е. могут быть реализованы по итогам обработки массивов эмпирических данных. Это вполне согласуется с теорией педагогических измерений, в соответствии с которой «тесты могут быть качественными и давать достоверные результаты только в том случае, если они предварительно апробированы на типичных выборках испытуемых и показывают соответствие заложенным при разработке взаимосвязанным свойствам надежности и валидности» [1, 36, 59, 60]. Однако для разработчика ТОЗУО важно оценить потенциальную достоверность оценивания учебных достижений обучающихся априори, на основе прогнозистических методик, базирующихся на математических моделях формирования оценок и расчета показателей, способных характеризовать априорную достоверность результатов тестирования.

Востребованность методик априорных оценок показателей качества ТОЗУО на этапе его конструирования обусловлена:

отсутствием в распоряжении разработчика теста достаточного для корректного применения статистических методов объема выборки тестируемых;

ограниченностью времени разработки теста τ_p , его апробации τ_a и собственно самого контроля учебных достижений обучающихся τ_k значениями $\Delta T_{\text{и}}$:

$$\tau_p + \tau_a + \tau_k < \Delta T_{\text{и}}; \quad (2.8)$$

необходимостью априорного выбора оптимальных решений в отношении структуры теста (распределения ТЗ по их видам и уровням трудности), соотношения объемов единичного теста и базы заданий, из которой он формируется, а также алгоритмов подсчета баллов и критериев оценок;

рисками снижения показателей надежности теста уже в процессе или по окончании его апробации из-за того, что тестируемым становится известной часть ТЗ, или существенным увеличением общей трудоемкости разработки теста при исключении применявшихся на этапе его апробации ТЗ из их базы и их замены на эквивалентные (такого рода замена при отсутствии в распоряжении разработчика теста ТЗ эквивалентного содержания и калиброванной (заранее определенной экспериментально) трудности может поставить под сомнение в целом корректность оценки качества теста, поскольку степень эквивалентности (по содержанию, трудности, дифференцирующей способности и др.) исключаемых из базы и вводимых вместо них, как правило, вновь созданных ТЗ определяется разработчиком теста субъективно.

Для оценки и априорной (потенциальной, прогнозируемой) и апостериорной (эмпирической) достоверности теста в педагогической литературе используются такие характеристики, как эффективность, пригодность, валидность и надежность [20, 36, 59].

Руководствуясь стандартизированным определением эффективности [61], следует рассматривать эффективность теста как отношение некоего количественного показателя, характеризующего достигнутый результат, к количественному показателю, указывающему на трудоемкость или стоимость получения данного результата. При вычислении знаменателя относительно несложно учесть (в часах или стоимости их оплаты) трудозатраты на разработку, апробацию, коррекцию теста, проведение контроля знаний и умений обучающихся, анализ и представление полученных результатов, а также стоимость приобретения и лицензионной поддержки программного обеспечения, эксплуатации оборудования и т. п., однако при вычислении числителя возникают трудности, вызванные неопределенностью достигаемого или желаемого приращения каких-либо количественных показателей характеристик теста, определяющих его качество [26]. Как показывает

анализ публикаций по вопросам качества дидактических тестов, эффективность теста обычно отождествляется авторами с его результативностью, т. е. действенностью – мерой достижения поставленной цели или планируемого результата. Поскольку результатом тестирования чаще всего является получение достоверных, соответствующих реальным знаниям и умениям обучающихся оценок их учебных достижений, такое отождествление приводит к тому, что чаще всего под эффективностью теста понимается именно достоверность оценивания им учебных достижений [26].

Одним из свойств теста, позволяющим определить его пригодность, является валидность. Проблемой применения термина «валидность» в качестве свойства ТОЗУО является то, что это понятие, заимствованное педагогической квалиметрией из смежных отраслей знаний (психологии, социологии), само по себе не подразумевает некоторое поддающееся измерению качество или характеристику.

Проблемой оценки валидности является необходимость обоснования одного или нескольких количественных показателей, ее характеризующих, и формирования соответствующих оценочных шкал с разработкой методик их применения. Без решения данной проблемы валидность ТОЗУО может характеризоваться лишь двумя достоверно определяемыми априори (например, методом экспертных оценок) или апостериори (по корреляции оценок теста с оценками, полученными заведомо более достоверными способами, их вариации (дисперсии) и др.) значениями – «валиден» или «не валиден», что существенно сужает ее диагностические возможности как характеристики качества тестов.

Первые педагогические тесты для проверки способностей к обучению или знаний и умений, в особенности тесты с начислением баллов, появились уже после того, как было сформулировано понятие валидности применительно к измерениям физических (сила, выносливость) или психических качеств человека [1, 15]. Вопрос валидности методики измерений возникает при определении величин посредством косвенных измерений, результаты которых связаны с определяемой величиной неявным или неоднозначным образом (например, определение степени ожирения человека по его росту и массе тела). Ее обоснование опирается либо на большой массив эмпирических данных, либо на экспертные оценки. В то же время оценка знаний и умений, основанная на балльной системе, нуждается в валидации только в отношении оценочной шкалы и критериев дифференцирования итоговых оценок теста, что наряду со структурой ТОЗУО, уров-

нем трудности заданий и алгоритмом подсчета баллов регулируется непосредственно разработчиком теста исходя из дидактических целей и частных задач диагностирования уровня учебных достижений обучающихся.

Длительное время педагогические измерения с помощью тестов были направлены на оценку интеллекта, общих или специальных знаний, выявление наклонностей и предпочтений в изучении тех или иных предметных областей, но не имели целью оценку результатов обучения. Так, экспериментальные и научные работы по созданию тестов для оценки знаний студентов стали предметом интереса педагогической общественности вузов СССР лишь в конце 60-х – начале 70-х гг. XX в., а первое признание тестирования как способа проверки уровня знаний, причем только в системе повышения квалификации, в СССР официально оформилось в декабре 1988 г. (постановление Госкомобразования СССР, Госкомтруда СССР, ВЦСПС от 23 декабря 1988 г. № 544/649/12-7 «Об утверждении Типового положения об учебных заведениях (подразделениях) системы повышения квалификации и переподготовки руководящих работников и специалистов народного хозяйства»). Привнесение термина «валидность» из психологической квалитметрии, где верификация результатов тестирования, как правило, осложняется отсутствием эталонных измерений оцениваемых качеств и, вследствие отсутствия или невысокой достоверности оценок, получаемых независимыми способами, зачастую осуществляется повторными измерениями в формате тестирования, что не исключает систематической ошибки измерений, в область оценки учебных достижений с результативно совершенствуемыми не одно столетие «классическими» формами оценки знаний и умений обучающихся сделало его применение в качестве характеристики достоверности оценок или свойства качества теста сопряженным с проблемами методологического характера.

При определении валидности оперируют либо абстрактными понятиями (например, «свойство истинности, правильности, соответствия реальности» или «степень, в которой тест измеряет то, что он должен измерять»), либо лишь косвенно упоминают о возможности ее количественного описания («комплексная характеристика методики (теста), включающая сведения об области исследуемых явлений и репрезентативности диагностической процедуры по отношению к ним»), либо методологически некорректно соотносят ее с точностью или достоверностью результатов тестирования («степень, в которой результаты исследования, системы измерений или статистики являются

точными или представляют то, что были предназначены представлять», «точность результатов теста или исследовательской работы; т. е. тот предел, до которого тест или исследование в действительности измеряет или показывает то, что надо измерить и показать», «достоверность, свидетельствует о степени правильности, истинности представленных данных, призванных подтвердить гипотезу»), которые в данном случае являются неопределенными понятиями как минимум того же уровня в терминологической иерархии, что и определяемый через них термин [26].

В [31] утверждается, что валидность теста «...зависит от качества заданий, их числа, степени полноты и глубины охвата содержания учебной дисциплины в заданиях теста; баланса и распределения заданий по трудности; метода отбора заданий из общего банка, от интерпретации результатов тестирования; организации сбора данных, отбора выборки испытуемых». Однако ни один из элементов данного перечня, ни их совокупность не могут быть рассмотрены в качестве меры, характеризующей степень валидности.

Для конкретизации данного понятия, столь неоднозначно трактуемого различными исследователями, в [36, 44, 45, 62–65] введены частные понятия валидности: «содержательная (в том числе внешняя)», «критериальная (статистическая, прогностическая)» «конструктивная (конструктивная)», «эмпирическая», «диагностическая (конкурентная) и прогностическая», «очевидная и концептуальная», «функциональная и критериальная», что согласуется с утверждениями исследователей о том, что «в понятие валидности входит самая разнообразная информация о тесте». Дефиниции перечисленных выше частных понятий зачастую содержат в себе другие неопределенные термины, например «конструктивная валидность, определяющаяся подбором точных и нужных для данного теста заданий, включает в себя понятие дифференцирующей силы вопроса, которая позволяет отобрать лучшие задания для окончательного варианта теста» [37]. Ни сами определения, ни пояснения к ним не содержат указания не только на способы, но и на возможности количественного описания каких-либо характеризующих их показателей, например «конструктивная валидность используется при сложности или невозможности подобрать адекватные критерии валидации. При этом используется комплекс характеристик, свидетельствующих о теоретической обоснованности методики, соответствии полученных с помощью теста результатов теоретическим ожиданиям и закономерностям» [1].

Наиболее поддающимся количественному описанию из перечисленных терминов является лишь эмпирическая валидность – «независимый показатель, в котором используются экспертные оценки и характеристики теста, данные специалистами», подразумевающая способ получения оценок – экспертный опрос, опирающийся на стандартизированные методики получения и обработки экспертных данных [6, 40, 42, 67], однако не определяющая совокупность оцениваемых показателей и критерии валидности теста на основании полученных оценок.

Наиболее близкий подход к представлению валидности количественными характеристиками сделан в [68]: «Валидность определяется посредством корреляции результатов тестирования с проявлением данного качества на практике. Валидность конкретных тестов может проверяться с помощью использования других, практически доказавших свою добротность методик. Мерой валидности служит коэффициент корреляции теста».

Применительно к тесту коэффициент корреляции должен означать корреляцию данных, полученных в разных условиях (например, для разных выборок тестируемых) по одному и тому же тесту, а в предлагаемом выше способе количественной оценки валидности корректнее говорить о коэффициенте корреляции статистических данных, полученных несколькими способами, включая рассматриваемый тест. Однако более очевидным представляется характеризовать валидность величиной отклонения результатов теста от неких заведомо достоверных данных, мерой которого является, например, дисперсия, вычисление которой как нормирующей величины сопровождает расчеты различных коэффициентов корреляции [69].

Приведенное в [40] определение: «количественной мерой критериальной валидности служат коэффициенты ранговой и бисериальной корреляции между показателями теста и критериальной мерой, задаваемой при конструировании теста» – в целом позволяет разработать методику оценки качества теста как меры соответствия проверяемых знаний и умений обучающихся каким-либо образцам, например требованиям образовательных стандартов, предполагаемым к достижению результатам обучения по образовательным программам и др., однако вопрос достоверности (в особенности априорной, потенциальной) получаемых оценок, т. е. их соответствия реальным знаниям и умениям обучающихся, остается за пределами определяемых таким образом количественных показателей.

Из анализа отечественных и зарубежных источников можно сделать вывод о том, что до настоящего времени не предложено приемлемой математической модели валидности ТОЗУО. Методики количественного или качественного описания параметров, характеризующих валидность тестов как показатель их качества или достоверности, включают в себя экспертные оценки или полностью на них основываются. Идея применения корреляционных методов оставляет открытым вопрос о выборе «критериальной меры» или множества оценок учебных достижений обучающихся для сравнения с результатами тестирования, и данный выбор в значительной мере определяется субъективно.

Не меньше сложностей в интерпретации и измерении надежности ТОЗУО. Практически все приводимые в педагогической и психологической литературе определения надежности теста объединяет отождествление данного термина с воспроизводимостью результатов тестирования, в частности «надежность теста понимается как способность давать одни и те же результаты при его применении к одинаковым выборкам тестируемых и характеризуется устойчивостью результатов тестирования» [68]. По способам получения сравниваемых результатов различают:

- тест-ретестовую надежность, характеризующуюся корреляционным сравнением результатов многократного выполнения одного теста;

- разделенную надежность, определяемую путем сравнения результатов выполнения двух отдельных частей теста;

- эквивалентную надежность, основанную на сравнении результатов решения испытуемым теста и его альтернативного варианта [70].

Хорошей воспроизводимостью могут обладать и недостоверные результаты, обусловленные, например, систематической ошибкой теста, поэтому в определении надежности теста иногда содержится упоминание о его точности как о второстепенном по сравнению с воспроизводимостью результатов свойстве. Так, «надежность теста отражает точность и устойчивость результатов тестирования к воздействию посторонних случайных факторов. Тест называется надежным, если он дает одни и те же (или очень близкие) показатели для каждого испытуемого при повторном тестировании» [40]. В [1] надежность рассматривается как характеристика воспроизведения в тесте базы ТЗ и одновременно точности и воспроизводимости результатов тестирования: «Надежность теста – показатель точности и устойчивости результатов измерения с помощью теста при его много-

кратном применении. Характеризует степень адекватности отражения тестом соответствующей генеральной совокупности заданий». Таким образом, точность и воспроизводимость результатов оценивания учебных достижений рассматриваются как независимые проявления надежности теста, однако ни один из авторов не задается вопросом: «Могут ли не воспроизводиться результаты оценивания учебных достижений, если тест обеспечивает их точность?».

В [52] отмечается, что расчет надежности достаточно сложен, и для практических целей рекомендуется «повторное тестирование испытуемых в одинаковых условиях по одним и тем же тестам» с последующим корреляционным анализом результатов и коэффициентом корреляции в качестве меры надежности.

Корреляционный подход к расчету показателей надежности теста не решает проблему ее априорной оценки. Недостатком корреляционных методов оценки надежности теста является то, что при рассмотрении тест-ретестовой надежности встает вопрос о корректности выбора множеств оценок для расчета коэффициента корреляции и корректности сопоставления результатов, если в повторных тестах полностью или частично повторяются уже известные испытуемым вопросы, а время тестирования не меняется. В других интерпретациях надежности (разделенной, эквивалентной) для обеспечения корректности результатов корреляционного анализа исходных и повторных данных требуется безукоризненное обоснование тождественности разделенных частей теста или первого и альтернативного тестов.

В [40] указывается другая мера надежности теста: «надежность связана с понятием стандартной ошибки измерения: чем выше надежность, тем меньше стандартная ошибка измерения». Однако без указания величины, относительно которой вычисляется стандартная ошибка, при использовании типовых методик обработки статистических данных, в которых стандартная ошибка рассчитывается относительно среднего значения, мера точности оказывается тождественной мере воспроизводимости.

Обратим внимание на то, что стандартная ошибка (среднеквадратическое отклонение) – квадратный корень из дисперсии, которая характеризует корреляцию результатов тестирования, т. е. валидность теста. Таким образом, валидность и надежность теста количественно оцениваются тождественными характеристиками, что так же делает их тождественными друг другу либо соотносит их как общее и частное.

Дисперсия оценок характеризует и дифференцирующую способность теста как необходимое условие достоверности оценивания

учебных достижений. Дифференцирующая способность подразумевает включение в тест большого количества ТЗ различного уровня трудности, что приводит к повышению содержательной валидности [36, 53, 71]. Однако при рассмотрении надежности исключительно как меры воспроизводимости результатов тестирования делается парадоксальный вывод о том, что при повышении его дифференцирующей способности и валидности теста снижается его надежность, а оптимальность надежности и валидности теста может быть достигнута путем подбора заданий сбалансированной трудности [52]. Парадоксальность вывода заключается в том, что валидность и надежность рассматриваются как основные характеристики качества теста, обеспечивающие достоверность оценок учебных достижений обучающихся, а ряд исследователей среди количественных показателей надежности и валидности отмечает близкие по смыслу величины. В данном аспекте корректным является утверждение о том, что чем выше погрешности оценок теста, количественно характеризующие надежность, тем ниже тестовая эффективность [60, 72].

По нашему мнению, основной причиной неизжитых терминологических противоречий в области оценки качества тестов является использование (в том числе вторичное заимствование, например из психологии) педагогической квалиметрией с собственной, зачастую изначально сугубо частной трактовкой на основе формальных лингвистических подходов или личного понимания учеными устоявшихся терминов из технических, естественнонаучных и математических дисциплин. Вторым по значимости фактором терминологических противоречий, на наш взгляд, является буквальный перенос понятий и определений теоретической тестологии в практику КУДО без учета специфики оцениваемых объектов, что породило несогласованность теории и практики в организации проверки знаний и умений обучающихся и интерпретации ее результатов. Проблемой методологического характера является несогласованное расширение терминологической базы современных педагогических исследований в области оценки качества ТОЗУО, выражающееся в терминологических новациях, противоречащих понятийному аппарату смежных наук, наделении устоявшихся понятий новыми признаками и свойствами, несистемном введении дополнительных частных терминов, поглощаемых по смыслу базовыми понятиями. Все это приводит к множественности и некорректности трактовок базовых терминов, подмене понятий и разрушению единства и общности научной терминологии.

Проблема корректной оценки качества тестов для контроля знаний имеет место в аспекте как терминологии, так и математического аппарата. Отсутствие корректных математических моделей ТОЗУО сказывается на возможности априорных оценок качества тестов и точности прогнозирования результатов тестирования. До некоторой степени прогноз успешности решения теста может быть получен на основе статистического метода (модели) Раша [8, 9, 23, 36, 50, 71, 73, 74] как вероятность решения ТЗ определенной трудности обучающимся, если известны результаты решения данного теста (**A**) группой других обучающихся и относительный уровень подготовки испытуемого, определяемый по результатам решения им тестов **B** аналогичной или калиброванной по одной шкале с **A** трудности (в том числе из других предметных или тематических областей). Применение данного метода дает наиболее достоверные результаты для тестов, составленных из ТЗ примерно одинаковой трудности, т. е. обладающих минимальной дифференцирующей способностью. Как любая статистическая модель, модель Раша опирается на большие объемы эмпирических данных, получение которых невозможно, например, преподавателем в процессе разработки им ФОС к учебной дисциплине. Модель в явном виде формально оперирует двумя величинами: уровнем подготовки группы испытуемых (испытуемого) U_1 и уровнем трудности T_A теста **A**, связанных между собой логистической функцией. Однако величина U_1 , в свою очередь, определяется по результатам решения данной группы теста **B** известной трудности T_B (частный случай: $T_B = T_A$), а величина T_A (или T_B) – по результатам решения теста **A** (или **B**) второй группы испытуемых с уровнем подготовки U_2 , т. е. неявно в модели Раша присутствуют минимум четыре определяемых статистической обработкой эмпирических данных величины – координаты двух точек логистической кривой: для нахождения неизвестной величины должны быть известны три другие. Недостаточный объем эмпирических данных (недостаточная репрезентативность выборок испытуемых для определения уровня трудности теста) или использование экспертных оценок может привести к большой дисперсии оценок искомых величин.

Корректность прогноза успешности решения ТОЗУО и достоверности получаемых на его основе оценок в значительной мере определяется корректностью модели испытуемого. Наиболее популярные модели испытуемого представляются некоторым интегральным уровнем подготовки, который с убывающей вероятностью позволяет решать задания возрастающей трудности [23, 35, 36, 73, 75].

На них основаны методики адаптивного тестирования, согласование уровней трудности тестов и прогнозирования результатов тестирования по методу Раша. Такие модели не учитывают содержательных лакун в знаниях испытуемых (в особенности на определенном этапе обучения), перекосы в соотношении «знания – умения» (хорошее знание теории и неумение решать задачи некоторой группой обучающихся или хорошие навыки решения задач при нехватке теоретических знаний у другой группы обучающихся из той же выборки тестируемых) и другие особенности сформированности знаний и умений, обусловленные как индивидуальными особенностями обучающихся (уровнем базовой подготовки, способностями к тому или иному виду учебной деятельности), так и результативностью педагогической деятельности преподавателей. Учет данных особенностей приводит не только к дополнительному увеличению дисперсии упомянутых выше априорных оценок качества тестов и результатов их применения, но и к расходимости таких оценок, что ограничивает использование подобных моделей узким кругом теоретических и прикладных задач.

В [26] нами обоснован подход к описанию качества ТОЗУО на основе вывода о тождественности алгоритма оценивания учебных достижений обучающихся системами тестирования алгоритмам работы автоматизированных измерительных систем, широко применяемых в технических системах контроля качества, управления объектами и технологическими процессами, безопасности, телеметрии и др., что позволяет применять методы математического моделирования функционирования и оценки качества, методологию и математический аппарат теории надежности технических систем и устройств данного класса для анализа качества ТОУЗО.

Анализ признаков сходства различных автоматизированных измерительных систем с КТОЗУО позволяет сделать вывод о том, что наиболее близким по предназначению, структуре (функциональной схеме) и алгоритму функционирования (принципу действия) к КТОЗУО является сортировочный (отбраковочный) автомат: КТОЗУО осуществляет автоматическую отбраковку (при двухбалльной оценочной шкале «зачет – незачет») или сортировку (при многобалльной оценочной шкале) знаний и умений испытуемых в соответствии с заданными критериями. Одной из распространенных реализаций такого автомата является система контроля и управления доступом, которая может выполнять функции допуска и недопуска на объ-

ект либо разграничения (определения уровня) доступа по вводимому паролю (цифровому коду) или биометрическим данным.

Согласно ГОСТ 27.002-89 надежность – «свойство объекта сохранять во времени в установленных пределах значения всех параметров, характеризующих способность выполнять требуемые функции в заданных режимах и условиях применения, технического обслуживания, хранения и транспортирования».

Очевидно, применительно к КТОЗУО надежность должна означать его способность в течение заданного времени правильно (с допустимой погрешностью) соотносить знания испытуемых с принятой оценочной шкалой.

Сопоставляя между собой понятия «надежность» и «достоверность» применительно к оцениванию учебных достижений, следует рассматривать надежность ТОЗУО как долгосрочную достоверность оценивания им учебных достижений, свойство сохранения достигнутых текущих показателей достоверности оценивания учебных достижений в некоторой временной (долгосрочной, среднесрочной) перспективе.

Надежность технических систем (изделий) оценивают по четырем показателям – безотказности, долговечности, ремонтпригодности и сохраняемости.

Безотказность – свойство объекта выполнять заданные функции непрерывно в течение некоторого времени или некоторой наработки.

Применительно к КТОЗУО безотказность – это свойство безошибочно соотносить знания испытуемых с принятой оценочной шкалой в течение некоторого времени или (и) некоторого количества тестирований.

Долговечность – свойство объекта выполнять заданные функции при установленной системе поддержания его работоспособности до наступления предельного состояния, при котором дальнейшая эксплуатация объекта недопустима или нецелесообразна.

Работоспособность КТОЗУО может поддерживаться путем изменения объема и содержания базы заданий и базы знаний, порогов и диапазонов оценок, т. е. залогом долговечности КТОЗУО является его регулярное обновление. Долговечность представляет собой комплексный параметр, зависящий от ремонтпригодности и сохраняемости объекта.

Ремонтпригодность для КТОЗУО может интерпретироваться как обеспечение возможности корректировки алгоритмов оценивания

знаний, объема и содержания базы заданий и базы знаний, а *сохраняемость* – как свойство с течением времени без использования сохранять актуальность без корректировки [26].

Из четырех приведенных характеристик лишь безотказность может быть пригодной для количественных оценок надежности КТОЗУО. Безотказность является также одной из четырех составляющих функциональной надежности программных продуктов наряду с работоспособностью, безопасностью и защищенностью [76]. Таким образом, безотказность представляет собой универсальный показатель надежности КТОЗУО, общий для его различных реализаций (аппаратных, программных, программно-аппаратных, текстовых документов).

Исходя из положений теории надежности [77, 78], безотказность КТОЗУО определяется вероятностью его работы без отказа в течение определенного периода времени.

Отказ ТОЗУО проявляется в совокупности ошибок первого и второго рода.

Ошибка первого рода для ответа на один вопрос выглядит так: испытуемый знает правильный ответ, но система расценивает его как неправильный. Ошибки первого рода приводят к занижению оценок КУДО.

Причинами ошибок первого рода могут быть:

- 1) ошибка отбора содержания ТЗ (включение в него неизученных вопросов или вопросов из других тем, разделов, предметных областей, плохо известных испытуемым);
- 2) некорректная формулировка вопроса и (или) вариантов ответов, приводящая к неоднозначному пониманию их смысла испытуемыми;
- 3) ошибка сопоставления вопроса и ответа, когда правильный ответ считается неправильным или правильный ответ отсутствует (из-за невнимательности или низкой квалификации составителя);
- 4) ошибка алгоритма начисления баллов за решенное (в том числе частично) ТЗ;
- 5) ошибка программирования теста;
- 6) ошибка программной оболочки теста;
- 7) ошибка ввода ответа тестируемым из-за невнимательности или неправильного использования интерфейса или бланка.

Ошибка второго рода для ответа на один вопрос проявляется в том, что система расценивает неправильный ответ испытуемого как

правильный. Ошибки второго рода приводят к завышению оценок учебных достижений обучающихся по итогам тестирования.

Причины ошибок второго рода:

- 1) ошибка отбора содержания теста (включение в него вопросов на общую эрудицию или вопросов из других, хорошо знакомых обучающимся тем, разделов, предметных областей);
- 2) некорректная формулировка вопроса, содержащая в себе подсказку или уменьшающая количество вариантов выбора правильного ответа;
- 3) ошибка сопоставления вопроса и ответа, когда неправильный ответ считается правильным (из-за невнимательности или низкой квалификации составителя);
- 4) ошибка алгоритма начисления балла за частично верное решение ТЗ;
- 5) ошибка программирования теста;
- 6) ошибка программной оболочки теста;
- 7) угадывание тестируемым правильного ответа;
- 8) ошибочный ввод тестируемым правильного ответа при выборе неправильного;
- 9) запоминание тестируемым правильного ответа без знания (из опыта прошлого тестирования);
- 10) подсказка тестируемому правильного ответа в ходе или накануне проведения теста;
- 11) использование тестируемым разного рода шпаргалок [26].

Ошибки и первого, и второго рода могут носить как систематический, так и случайный характер. Причины систематических ошибок поддаются выявлению и устранению путем корректировки тестов, а также организационными мерами. Ошибки случайного характера могут быть оценены и учтены путем статистической обработки результатов многократных тестирований. Как правило, их вклад в общую величину ошибки КТОЗУО относительно невелик и не оказывает существенного влияния на результат. Ряд причин ошибок, например угадывание правильного ответа, носит комбинированный характер, однако явное наличие в них случайной составляющей делает оправданным применение для оценки их величины аппарата теории вероятностей и математической статистики [4].

Ошибки КТОЗУО можно классифицировать по аналогии с отказами технических систем по ГОСТ 27.002–89. Так, по возможности дальнейшего использования или восстановления объекта отказы разделяют на *полные* и *частичные*. Применительно к КТОЗУО частич-

ный отказ следует понимать как «взлом» части теста настолько, что знание тестируемым ответов на отдельные вопросы приводит к завышенной оценке (ошибка второго рода), или как утрату актуальности частью базы знаний, ее несоответствие современным учениям, новейшим открытиям, достижениям техники, дидактики и т. п., в частности содержанию обучения (ошибка второго рода), что особенно характерно для общественно-научных и юридических учебных дисциплин. Характерные примеры: принятие поправок в Конституцию РФ на Всероссийском референдуме, внесение изменений в законодательные акты, исключение Плутона из числа планет Солнечной системы. При групповом тестировании частичный отказ означает завышение оценок части контингента тестируемых вследствие знания ими верных ответов без освоения соответствующих дидактических единиц в целом. Причиной частичного отказа могут быть и незначительные ошибки в алгоритме тестирования, например процедуре начисления баллов и задании оценочной шкалы или порога положительной оценки. Немаловажным фактором возникновения частичных отказов КТОЗУО при групповом тестировании, в особенности при проверке остаточных знаний [30], играет низкая мотивация отдельных обучающихся.

Полный отказ теста следует трактовать как очевидную, не согласующуюся с результатами других видов контроля знаний и умений тестируемых недостоверность оценок, воспроизводимую при повторном тестировании одного испытуемого или проявившуюся для большинства тестируемых в группе.

Помимо недостоверной оценки учебных достижений, отказы приводят к ошибкам анализа результатов тестирования на предмет успешности освоения обучаемыми содержания обучения, уровня их остаточных знаний, эффективности методики обучения и применения отдельных педагогических приемов при проведении педагогических экспериментов и др.

По динамике проявления отказов или изменению основного свойства объекта различают *постепенные* и *внезапные* отказы. Для КТОЗУО постепенный отказ характеризует его деградацию, механизм которой определяется накоплением ошибок (как правило, второго рода) и проявляется в постепенном (от одной группы тестируемых к другой) увеличении дисперсии отклонения оценок теста от оценок других видов проверки знаний и умений или оценок предыдущих тестов данного контингента обучающихся. Внезапный отказ теста выглядит как резкое, проявляющееся для одного сеанса тестирования несоответствие полученных и прогнозируемых на основе имеющихся

эмпирических данных по другим видам контроля оценок обучающихся. Наиболее часто ситуация, которая может трактоваться как внезапный отказ КТОЗУО, имеет место при контроле остаточных знаний, и ее причиной могут быть как несоответствие уровня заданий уровню остаточных знаний (слишком легкие или слишком трудные ТЗ, приводящие к ошибкам соответственно второго и первого рода), несогласованность содержания тестов со структурой остаточных знаний, так и предварительное ознакомление обучающихся с заданиями теста и (или) ответами на них.

Количественными характеристиками отказов являются вероятность отказа P_0 и интенсивность отказов $\lambda_{отк}(t)$. Интенсивность отказов КТОЗУО может определяться как отношение количества неверных оценок, полученных в результате КУДО, к общему количеству таких оценок за некоторый период времени. График $\lambda_{отк}(t)$ представляется корытообразной кривой [77, 78], на которой выделяют три характерных участка (рис. 2.1).

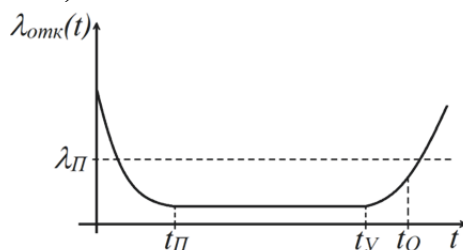


Рис. 2.1. Типичный график зависимости интенсивности отказов от времени

Первый участок приходится на временной интервал $[0; t_p]$, где t_p – время приработки теста, за который интенсивность отказов КТОЗУО (ошибочных оценок) снижается от некоторого начального значения до величины, близкой к постоянной. Наличие данного участка обусловлено ошибками первого и второго рода при составлении теста и последующей работой над их устранением. Ненулевые значения $\lambda_{отк}(t)$ на втором интервале $[t_p; t_y]$, где t_y – время устаревания теста, обусловлены наличием неустраненных ошибок КТОЗУО и случайными причинами возникновения ошибок первого и второго рода. На третьем временном интервале $[t_y; t_o]$, где t_o – время окончания использования теста в системе контроля знаний обучающихся, рост интенсивности отказов, как правило, обусловлен тем, что тест становится хорошо знакомым обучающимся и количество правильных ответов, которые им известны без реальных знаний, соответствующих ДЕСО учебной дисциплины, увеличивается от тестирования к тестированию. Отсутствие заинтересованности разработчика по корректи-

ровке теста как меры по снижению роста $\lambda_{\text{отк}}(t)$ может быть вызвано, например, разработкой нового теста или прекращением действия срока договора на доработку соответствующей компьютерной контролирующей программы.

Рассмотренные временные интервалы и отнесенные к ним участки кривой $\lambda_{\text{отк}}(t)$ называются соответственно периодами (участками) приработки, нормальной эксплуатации и старения [77, 78].

Приемлемый уровень $\lambda_{\text{отк}}(t)$, характеризующийся величиной $\lambda_{\text{п}}$, выбирается организаторами проверки знаний исходя из задач контроля. Начальное значение $\lambda_{\text{отк}}(0)$ может оказаться меньше $\lambda_{\text{п}}$, и в этом случае корректировки теста не требуется, поэтому начальный участок спада $\lambda_{\text{отк}}(t)$ будет отсутствовать. На графике $\lambda_{\text{отк}}(t)$ может отсутствовать и явно выраженный участок старения ввиду того, что эксплуатация КТОЗУО может быть прекращена до того, как $\lambda_{\text{отк}}(t)$ превысит значение $\lambda_{\text{п}}$.

Вероятность отказа КТОЗУО, очевидно, равна сумме вероятностей ошибок первого и второго рода:

$$P_o = P_{o1} + P_{o2}. \quad (2.10)$$

При эксплуатации теста, в основном за период времени $[0; t_{\text{п}}]$, работчики выявляют и исправляют систематические ошибки первого рода (1–6) и частично ошибки второго рода (1–6), а также в течение всего периода эксплуатации принимают меры по устранению ошибок второго рода (10, 11).

Соответствующее снижение вероятности ошибок первого и второго рода:

$$P'_{o1} = P'_{\text{нач1}} \cdot e^{\frac{-t \cdot \alpha_{k1}}{\Delta t_k}}, \quad (2.11)$$

$$P'_{o2} = P'_{\text{нач2}} \cdot e^{\frac{-t \cdot \alpha_{k2}}{\Delta t_k}}, \quad (2.12)$$

где, $P'_{\text{нач1}}$, $P'_{\text{нач2}}$ – соответствующие начальные значения, Δt_k – эффективный период приработки теста, α_{k1} и α_{k2} – показатели эффективности устранения ошибок.

В результате деградации теста, проявляющейся в увеличении ошибок второго рода, интенсивность отказов теста возрастает. Вероятность ошибок второго рода, обусловленная факторами 7, 9:

$$P''_{o2} = P''_{\text{нач3}} \cdot e^{\gamma t}, \quad (2.13)$$

где γ – обобщенный коэффициент, зависящий от уровня запоминания, забывчивости и обучения испытуемого.

Следует отметить, что сумма убывающих (2.11), (2.12) и возрастающей (2.13) экспонент дает корытообразную кривую (рис. 2.1.).

Таким образом, корректная оценка качества ТОЗУО может основываться на теории надежности – отрасли научного знания, характеризующейся универсальностью, устоявшейся терминологией и математическим аппаратом, интегрированностью в научно-технические дисциплины, всесторонне и верифицируемой разнообразной и обширной эмпирической базой. Такой подход способен устранить имеющиеся противоречия в терминологии, моделях и интерпретации результатов тестирования. Термины теории надежности адекватно интерпретируются в формализованных и автоматизированных системах контроля знаний, что позволяет на основе ее математического аппарата разрабатывать различные модели и методики оценки качества тестов для проверки знаний обучающихся с учетом специфики реализуемых алгоритмов тестирования, соотнесения уровня знаний с уровнем трудности теста, динамических воздействий и др.

2.2. Общий подход к моделированию надежности теста для контроля знаний и умений обучающихся

В [26, 29, 75] нами обосновано применение надежности в качестве характеристики достоверности и основного показателя качества КТОЗУО. По аналогии с надежностью технических устройств надежность КТОЗУО означает его способность в течение заданного времени правильно (с допустимой погрешностью) соотносить знания испытуемых с принятой оценочной шкалой.

В процессе применения КТОЗУО систематически встает вопрос о достоверности результатов тестирования, связанный с адаптацией обучающихся к ТЗ, наиболее проявляющейся в заучивании ими правильных ответов без усвоения содержания изучаемого материала. Данное обстоятельство необходимо учитывать при анализе надежности КТОЗУО посредством построения модели испытуемого, которая должна соотносить текущий уровень знаний испытуемого и трудность ТЗ с вероятностью угадывания верных ответов, учитывать влияние на результаты тестирования процессов забывания и восстановления знаний.

Вероятность безотказной работы теста с учетом (2.10):

$$P_B = 1 - P_0 = 1 - (P_{01} + P_{02}). \quad (2.14)$$

Вероятность угадывания ответа на вопрос в ТЗВО, приводящего к ошибкам второго рода:

$$P_{\text{угад}} = \frac{1}{n - n_{\text{искл}}}, \quad (2.15)$$

где n – количество вариантов ответа, $n_{\text{искл}}$ – количество ответов, которые испытуемый исключает, опираясь на остаточные знания, общую эрудицию, знания из смежных предметных областей.

Подверженность ТЗВО ошибкам второго рода, порожденным угадыванием испытуемым правильных ответов, приводит к значительным ошибкам второго рода КТОЗУО, использующих ТЗ данного типа, и существенно повышает вероятность их отказа (незаслуженных положительных оценок по шкале «зачет – незачет» или завышенных оценок при их многоуровневой дифференцированной шкале).

Повышение вероятности безотказной работы КТОЗУО с ТЗВО достигается, как правило, за счет регулярного обновления базы ТЗ, из которой формируется тест: замены ТЗ или их редактирования (частичной замены вариантов ответа или изменения их формулировок), а также повышением порога положительной оценки или пересмотром в целом оценочной шкалы по мере деградации теста.

Снижение вероятности ошибки второго рода КТОЗУО может достигаться применением ТЗ с усложненной структурой (ТЗМВ, ТЗУС и др. [61]), а также усложнением процедур проверки результатов тестирования (например, на основе двухступенчатых алгоритмов, адаптивных алгоритмов и других реализаций обратных связей в процедурах подбора ТЗ в режиме реального времени и начисления баллов).

Величина $n_{\text{искл}}$ является функцией времени с определяемой индивидуальными способностями испытуемого и дидактическими условиями постоянной $\tau_{\text{искл}} \gg \tau_{\text{т}}$, где $\tau_{\text{т}}$ – обобщенное время тестирования, так как количество знаний, на которые опирается испытуемый при ответе, изменяется со временем в результате забывания ранее изученного материала и получения нового объема знаний (обучения), включая восстановление ранее забытых:

$$n_{\text{искл}}(t) = n_{\text{искл}}^* + n_{\text{искл}}^{\text{обуч}}(t) - n_{\text{искл}}^{\text{забыв}}(t), \quad (2.16)$$

где $n_{\text{искл}}^* \geq 0$ – некоторый незабываемый остаток.

Функция, характеризующая накопление и восстановления знаний, представляет собой сумму парциальных ступенчатых функций, величина которых $n_i^{\text{обуч}}$ отражает объем знаний, соотнесенный с ДЕСО, а времена «включения», в общем случае связанные функцио-

нальной зависимостью с $\tau_{\text{искл}}$: $t_i = f(\tau_{\text{искл}})$, – интенсивность его изучения и восстановления:

$$n_{\text{искл}}^{\text{обуч}}(t) = \sum_{i=1}^{v(t_{\text{обуч}})} n_i^{\text{обуч}} \sigma(t - t_i), \quad (2.17)$$

где $\sigma(t - t_i)$ – функция Хевисайда [79], а предел суммирования $v(t_{\text{обуч}})$ определяется временем обучения (подготовки к тестированию) $t_{\text{обуч}}$.

Забывание описывается экспоненциальной функцией [26, 57, 58]:

$$n_{\text{искл}}^{\text{забыв}}(t) = [n_0 \cdot e^{-\beta t}], \quad (2.18)$$

где β – показатель забывчивости, определяемый индивидуальными особенностями тестируемого (памятью, ассоциативным мышлением, владением методиками удержания в памяти информации) и системностью рассматриваемых знаний, $[x]$ – округление числа x в большую сторону.

С учетом выражений (2.16), (2.17), (2.18):

$$P_{\text{угад}}(t) = \left[n - n_{\text{искл}}^* - \sum_{i=1}^{v(t_{\text{обуч}})} n_i^{\text{обуч}} \sigma(t - t_i) + [n_0 \cdot e^{-\beta t}] \right]^{-1}. \quad (2.15a)$$

Отвечая на N вопросов теста, обучающийся дает $M_{\text{пр}}$ правильных ответов, часть из которых он знает, а часть угадывает:

$$M_{\text{пр}}(t) = M_{\text{зн}}(t) + \Delta M_{\text{уг}}(t). \quad (2.19)$$

Для получения положительной оценки необходимо ответить правильно на $M_{\text{зач}}$ вопросов, т. е. должно выполняться условие:

$$\frac{M_{\text{пр}}}{M_{\text{зач}}} \geq 1, \quad (2.20)$$

или

$$\frac{M_{\text{зн}}(t) + \Delta M_{\text{уг}}(t)}{M_{\text{зач}}} \geq 1. \quad (2.21a)$$

Отсюда необходимое условие безотказной работы теста:

$$\frac{M_{\text{зн}}(t)}{M_{\text{зач}}} \geq 1. \quad (2.21b).$$

В [26] приведены оценки надежности КТОЗУО, составленных из различных видов ТЗ закрытого типа: ТЗВО, ТЗМВ, ТЗУС, ТЗУП, ТЗГО, ТЗУД. Отличительной особенностью методики расчетов вероятности преодоления порога положительной оценки, интерпретируемой как вероятность отказа КТОЗУО, является учет уровня обученности испытуемых. Модель испытуемого характеризуется несколькими уровнями обученности в соответствии с усвоенным объемом контролируемого содержания обучения, что позволяет для каждого уровня

вычислить вероятность получения положительной оценки с учетом угадывания неизвестных ответов. Каждый уровень обученности подразумевает не только знание определенного процента верных ответов, но и частичное исключение дистракторов (для ТЗВО и ТЗМВ), частичное знание верных ответов, например одного из нескольких верных ответов в ТЗМВ или одного или двух парных соответствий в ТЗУП.

В ходе вычислений применялись различные алгоритмы назначения баллов за частично верные ответы на ТЗ с усложненной структурой, одновременно характеризующиеся едиными подходами и учетом особенностей структуры ТЗ. Для корректного сравнения величин ошибок второго рода при применении тестов на основе ТЗ одного вида использовалась двухбалльная шкала оценок («зачет – незачет») с одинаковым для всех анализируемых структур тестов и видов ТЗ уровнем удовлетворительной оценки – 80% от максимально возможного количества набранных баллов.

Сопоставление по единой методике показателей надежности тестов с ТЗ различных видов представляет интерес для оптимизации структуры тестов, включающей распределение проверяемых ДЕСО по видам ТЗ, последовательность и временные лимиты тестирования.

Разделение вопросов по уровням трудности позволяет вводить дополнительные критерии удовлетворительной оценки, тем самым повышая надежность теста за счет уменьшения вероятности ошибок второго рода. Если каждому уровню трудности (легкий, средней трудности, трудный) присвоить весовой коэффициент: $P_{\text{легк}}$, $P_{\text{ср}}$, $P_{\text{тр}}$, то условие успешного прохождения ТОЗУО:

$$N_{\text{легк}}^{\text{отв}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{отв}} \cdot P_{\text{ср}} + N_{\text{тр}}^{\text{отв}} \cdot P_{\text{тр}} \geq T_{\text{зач min}}, \quad (2.22)$$

где $T_{\text{зач min}}$ – минимальное достаточное для получения положительной оценки количество баллов, $N_{\text{легк}}^{\text{отв}}$, $N_{\text{ср}}^{\text{отв}}$, $N_{\text{тр}}^{\text{отв}}$ – количество правильных ответов на вопросы, отнесенные соответственно к категориям легких, средней трудности и трудных.

Нормальной для КУДО является ситуация, когда посредственно подготовленный испытуемый не отвечает на трудные вопросы и дает правильные ответы на часть вопросов средней трудности и большинство легких вопросов. Повышение уровня подготовки влечет за собой увеличение количества правильных ответов на вопросы во всех категориях, однако ситуация, когда при тестировании процент правильных ответов на трудные вопросы больше, чем процент безошибочно отвеченных легких вопросов, свидетельствует скорее о «взломе» те-

ста или использовании испытуемым подсказок. Разделение вопросов по уровням трудности позволяет контролировать корректность процедур тестирования и надежность самих тестов.

Поскольку на трудные вопросы может ответить меньшинство испытуемых, баллы за них носят премиальный характер, и набор $T_{зач\ min}$ должен быть обеспечен за счет ответов на вопросы средней трудности и легкие:

$$N_{\text{легк}}^{\text{отв}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{отв}} \cdot P_{\text{ср}} \geq T_{зач\ min}; \quad (2.22a)$$

$$N_{\text{легк}}^{\text{вопр}} \geq N_{\text{легк}}^{\text{отв}}; \quad N_{\text{ср}}^{\text{вопр}} \geq N_{\text{ср}}^{\text{отв}}.$$

Здесь $N_{\text{легк}}^{\text{вопр}}, N_{\text{ср}}^{\text{вопр}}$ – количество легких и средней трудности вопросов в ТОЗУО.

Справедливо полагать, что правильные ответы на все легкие и средней трудности вопросы должны гарантировать положительную оценку, в частности:

$$N_{\text{легк}}^{\text{вопр}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{вопр}} \cdot P_{\text{ср}} > T_{зач\ min}. \quad (2.23)$$

Отсюда величина достаточного для положительной оценки количества баллов может быть выражена через количество вопросов категорий легких и средней трудности:

$$N_{\text{легк}}^{\text{вопр}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{вопр}} \cdot P_{\text{ср}} = \frac{T_{зач\ min}}{k^*}. \quad (2.24)$$

Здесь $k^* < 1$ – варьируемый коэффициент, характеризующий уровень требований к знаниям испытуемых.

Дополнительными условиями критерия удовлетворительной оценки испытуемого, позволяющего контролировать наличие ошибок второго рода и безотказность работы теста в целом, могут быть требования ответов на определенное минимальное количество легких вопросов и вопросов средней трудности:

$$N_{\text{легк}}^{\text{отв}} \geq N_{\text{легк}\ min}^{\text{отв}} = N_{\text{легк}}^{\text{вопр}} \cdot k_{\text{легк}}^*; \quad (2.25a)$$

$$N_{\text{ср}}^{\text{отв}} \geq N_{\text{ср}\ min}^{\text{отв}} = N_{\text{ср}}^{\text{вопр}} \cdot k_{\text{ср}}^*. \quad (2.25b)$$

Здесь $k_{\text{легк}}^* < 1$ и $k_{\text{ср}}^* < 1$ – коэффициенты, выбираемые организатором КУДО с целью выявления ошибок второго рода.

Из предположения, что испытуемый отвечает на больший процент легких вопросов, чем вопросов средней трудности ($k_{\text{легк}}^* > k_{\text{ср}}^*$), набор количества баллов, превышающего $T_{зач\ min}$, при несоблюдении условия (2.20) может быть расценен как ошибка второго рода.

Если критерий удовлетворительной оценки дополнен требованием решения какого-то количества трудных вопросов ($k_{\text{тр}}^* > 0$), то во избежание противоречивых толкований результатов теста необходимо соблюдение условия:

$$N_{\text{легк}}^{\text{вопр}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{вопр}} \cdot P_{\text{ср}} < T_{\text{зач min}}. \quad (2.23a)$$

На рис. 2.2 приведен пример дерева событий прохождения теста с дополнительными скрытыми для тестируемых требованиями на решение определенного количества легких, средней трудности и трудных вопросов [29].

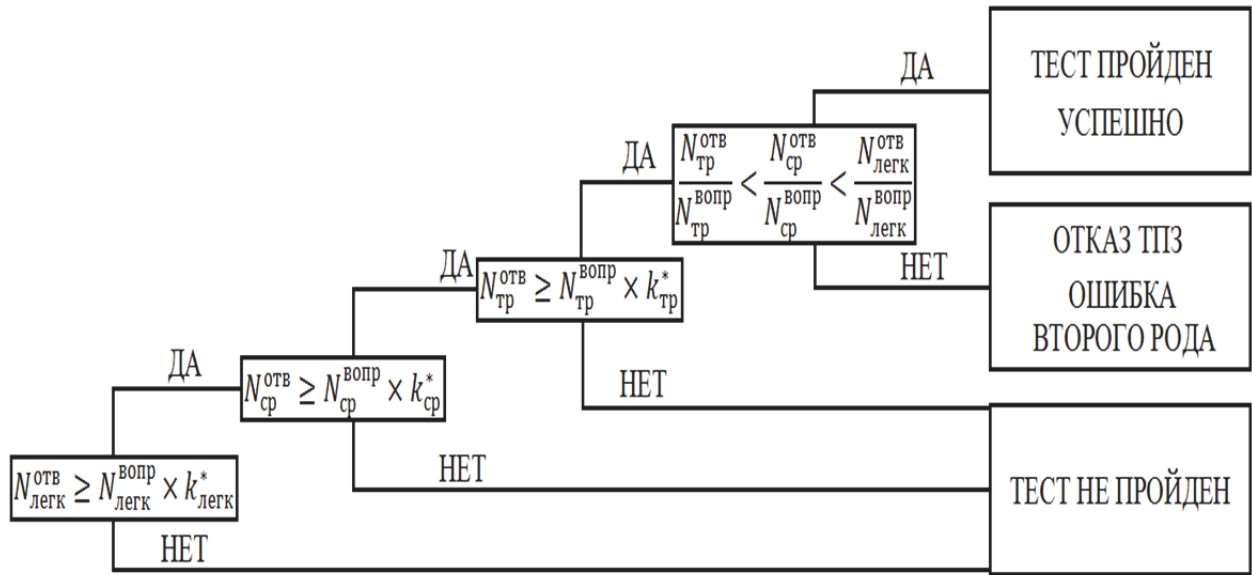


Рис. 2.2. Пример дерева событий прохождения теста

Условие (2.22) с учетом дополнительных требований можно привести к виду:

$$N_{\text{легк}}^{\text{отв}} \cdot P_{\text{легк}} \cdot \varphi_{\text{легк}} + N_{\text{ср}}^{\text{отв}} \cdot P_{\text{ср}} \cdot \varphi_{\text{ср}} + N_{\text{тр}}^{\text{отв}} \cdot P_{\text{тр}} \cdot \varphi_{\text{тр}} \geq T_{\text{зач min}}, \quad (2.22б)$$

где φ определяется следующим образом:

$$\varphi_{\text{легк}} = \begin{cases} 0, \text{ если } \frac{N_{\text{легк}}^{\text{отв}}}{N_{\text{легк}}^{\text{вопр}}} < k_{\text{легк}}^* \\ 1, \text{ если } \frac{N_{\text{легк}}^{\text{отв}}}{N_{\text{легк}}^{\text{вопр}}} \geq k_{\text{легк}}^* \end{cases};$$

$$\varphi_{\text{ср}} = \begin{cases} 0, \text{ если } \frac{N_{\text{ср}}^{\text{отв}}}{N_{\text{ср}}^{\text{вопр}}} < k_{\text{ср}}^* \\ 1, \text{ если } \frac{N_{\text{ср}}^{\text{отв}}}{N_{\text{ср}}^{\text{вопр}}} \geq k_{\text{ср}}^* \end{cases};$$

$$\varphi_{\text{тр}} = \begin{cases} 0, \text{ если } \frac{N_{\text{тр}}^{\text{отв}}}{N_{\text{тр}}^{\text{вопр}}} < k_{\text{тр}}^* \\ 1, \text{ если } \frac{N_{\text{тр}}^{\text{отв}}}{N_{\text{тр}}^{\text{вопр}}} \geq k_{\text{тр}}^* \end{cases}.$$

Оценивание учебных достижений и контроль надежности КТОЗУО могут осуществляться на основе экспертной модели, соотносящей уровни трудности ТЗ и подготовки испытуемых и основанной на трех экспертных гипотезах.

1. Испытуемый, претендующий на оценку не ниже «удовлетворительно», обязан правильно решить не менее $W_{LS} \cdot 100\%$ легких ТЗ, не менее $W_{MS} \cdot 100\%$ заданий средней трудности и не менее $W_{HS} \cdot 100\%$ ТЗ повышенной трудности. Значение W_{HS} может полагаться равным нулю.

2. Испытуемый, претендующий на оценку не ниже «хорошо», обязан правильно решить не менее $W_{LG} \cdot 100\%$ легких ТЗ, не менее $W_{MG} \cdot 100\%$ заданий средней трудности и не менее $W_{HG} \cdot 100\%$ трудных (повышенной трудности) ТЗ. При этом должны соблюдаться условия:

$$W_{LG} \geq W_{LS}; \quad (2.26a)$$

$$W_{MG} \geq W_{MS}; \quad (2.26b)$$

$$W_{HG} \geq W_{HS}; \quad (2.26b)$$

$$W_{LG} + W_{MG} + W_{HG} > W_{LS} + W_{MS} + W_{HS}. \quad (2.26g)$$

3. Испытуемый, претендующий на оценку «отлично», обязан правильно решить не менее $W_{LE} \cdot 100\%$ легких ТЗ, не менее $W_{ME} \cdot 100\%$ заданий средней трудности и не менее $W_{HE} \cdot 100\%$ ТЗ повышенной трудности. При этом должны соблюдаться условия:

$$W_{LE} \geq W_{LG}; \quad (2.27a)$$

$$W_{ME} \geq W_{MG}; \quad (2.27b)$$

$$W_{HE} \geq W_{HG}; \quad (2.27b)$$

$$W_{LE} + W_{ME} + W_{HE} > W_{LG} + W_{MG} + W_{HG}. \quad (2.27g)$$

Для задания значений W_{LS} , W_{MS} , W_{HS} , W_{LG} , W_{MG} , W_{HG} , W_{LE} , W_{ME} , W_{HE} может быть использована как статистическая модель Раша, так и экспертные оценки, учитывающие вид и содержательную трудность ТЗ, особенности начисления баллов за их частично верное решение. Инцидент надежности ТОЗУО заключается в выходе результатов испытуемого за пределы его модели, т. е. в нарушении условий (2.26) или (2.27).

2.3. Учет видов тестовых заданий при моделировании надежности теста для контроля знаний и умений обучающихся

Противоречивые требования к свойствам КТОЗУО можно свести к формулировке следующей задачи для разработчика теста: обеспечение требуемых показателей надежности минимальным количеством ТЗ. На практике на первоначальном этапе проектирования КТОЗУО обеспечение его надежности сводится в основном к реализации двух важнейших свойств: дифференцирующей способности и содержательной валидности. Одним из направлений решения данной задачи является усложнение структуры КТОЗУО следующими способами:

- усложнение структуры ТЗ;
- расширение видового многообразия ТЗ;
- усложнение алгоритма начисления баллов за решение ТЗ;
- формирование тестовых выборок из ТЗ различных видов, структуры и трудности;
- усложнение алгоритма перевода баллов в оценки.

Выбор ТЗ для повышения компактности КТОЗУО и, соответственно, оперативности тестирования при приемлемой достоверности оценок должен быть ориентирован на высокую степень устойчивости КТОЗУО к деградации (следует учитывать механическое запоминание правильных ответов при многократном индивидуальном или групповом тестировании) и угадыванию правильного ответа. Разнообразие видов ТЗ в тесте, включая ТЗ с усложненной структурой и ТЗ открытого типа [8, 27, 38], требует от испытуемых дополнительной мобилизации, умения быстро перестраивать мыслительные процессы и способствует повышению объективности КУДО, одновременно значительно расширяя творческий инструментарий преподавателей – разработчиков КТОЗУО.

Тестовое задание с множественным выбором верных ответов за счет возможности оценивать частично верные ответы обладает потенциально высокой дифференцирующей способностью. Примерные способы оценивания ТЗМВ [80]:

- 1) назначение одного балла за полностью правильно выполненное задание, т. е. за K верных ответов из K выбранных вариантов и ноль баллов за все другие возможные комбинации выбора из N предлагаемых вариантов;

2) зачет выбора верных ответов, например прибавление $1/K$ балла за каждый выбранный верный ответ из K возможных при условии невыбора неверного ответа;

3) штраф за выбор неверных ответов, например вычитание $1/(N - K)$ балла за каждый выбранный неверный ответ;

4) зачет выбора верных вариантов и штраф за выбор неверных вариантов ответов одновременно без дополнительных условий;

5) штраф за невыбор ни одного из вариантов ответа, например вычитание $1/(N - K)$ балла.

Каждому способу оценивания соответствует свое пороговое значение удовлетворительной оценки, которое может быть получено на основании модели испытуемого статистическими или экспертными методами [26]. Усложнение процедуры начисления баллов за частично верные ответы позволяет не только повысить дифференцирующую способность ТЗМВ, но и более детально учитывать ошибки первого и особенно второго рода и за счет этого как минимум в два раза сократить количество ТЗ, сохранив значение вероятности угадывания верных ответов на предельно низком уровне.

Среди ТЗУС нами рассмотрен наиболее простой вариант их реализации – ТЗ на установление парных соответствий между множествами понятий (например, «Обь, Лена, Печора, Северная Двина – Белое, Баренцево, Карское, Лаптевых», «индуктивность, проводимость, емкость, сопротивление – Сименс, Фарада, Ом, Генри»), реализуемых во многих программных средах для дистанционного обучения (в частности, в популярной СУО Moodle). При расчете вероятностей ошибок второго рода за счет угадывания верных ответов полагалось, что все соответствия взаимоисключающие, т. е. одному элементу первого множества соответствует только один элемент второго множества, и наоборот. Как и для ТЗМВ, основным объектом моделирования был тест из 20 ТЗ одного вида.

Согласно полученным результатам, даже относительно высокий уровень подготовки ($3/4$ от минимально достаточного количества $M_{\text{зач}}$ априори правильно решенных ТЗУС и $3/8$ от $M_{\text{зач}}$ наполовину известных решений ТЗУС) не гарантирует положительной оценки. При доле известных ответов $M_{\text{зн}} \leq 0,6 \cdot M_{\text{зач}}$ вероятность получения положительной оценки путем угадывания неизвестных ответов становится пренебрежимо малой. Для относительно высоких уровней подготовки испытуемых вероятность ошибки второго рода меньше, чем дает тест, составленный из ТЗМВ и наиболее простыми алгоритмами оценивания, и немногим больше, чем обеспечивает тот же тест

с усложненным алгоритмом оценивания, предусматривающим штрафы за выбор неверных вариантов ответов [80].

Тестовые задания на установление последовательности применяются разработчиками КТОЗУО сравнительно редко, что объясняется относительной сложностью конструирования ТЗ данного вида и сравнительно узкой спецификой их применения. Тем не менее для КУДО по дисциплинам, использующим символьные выражения, устойчивые лингвистические конструкции, унифицированный понятийный аппарат, наработан большой опыт разработки и применения такого специфического вида ТЗ. В целях простоты программной реализации проверочных алгоритмов изначально расставленные случайным образом элементы таких последовательностей нумеруются по порядку, и ответ на единичные ТЗ представляет собой последовательность цифр с измененным (верным) порядком следования элементов последовательности.

Пример 2.1: «is, task, it, difficult, not».

Пример 2.2: укажите последовательность операций в логическом выражении $a \rightarrow b \wedge c \leftrightarrow d \vee e$: конъюнкция, дизъюнкция, импликация, эквиваленция.

Вероятность угадывания правильной последовательности «вслепую» такая же, как для ТЗУС с парными соответствиями, однако в ряде последовательностей (как правило, формулировках определений, правилах) во избежание нескольких формально правильных вариантов ответа одна (обычно первая) или две позиции предопределяются логикой построения последовательности, являясь ее отправными или связующими элементами, что повышает вероятность угадывания верных ответов и определяет особенность моделирования надежности КТОЗУО, составленных из ТЗУП, включая модель испытуемого в части учета априори известных ему сведений [81].

Повышение надежности КТОЗУО и субтестов, составленных из ТЗУП, может быть достигнуто за счет увеличения количества позиций в ТЗУП и исключения последовательностей с фиксированными позициями элементов. Однако, как показывает наш опыт применения таких ТЗ, увеличение позиций последовательности до шести и более приводит к существенному, в два и более раза, по сравнению с ТЗУП, содержащих пять (из которых одна или две фиксированные) или четыре элемента последовательности, увеличению времени решения заданий. Усложнение ТЗУП за счет увеличения числа комбинаций элементов приводит к наибольшему среди рассмотренных нами видов ТЗ

увеличению вероятности ошибок первого рода, связанных как с ошибками при вводе ответа, так и с ошибками разработчиков КТОЗУО.

Сопоставление результатов моделирования надежности КТОЗУО, составленных из ТЗ с усложненной структурой, позволяет сделать следующие выводы.

1. Вероятность угадывания верного ответа на ТЗМВ, ТЗУС и ТЗУП существенно меньше, чем на ТЗВО, что позволяет формировать из ТЗ с усложненной структурой относительно компактные и надежные КТОЗУО.

2. Вероятности угадывания верного ответа для всех трех рассмотренных нами типов ТЗ с усложненной структурой близки, что позволяет комбинировать ТЗ данных видов в КТОЗУО, заменяя один другим, без необходимости повторных оценок надежности КТОЗУО.

3. Тестовые задания на установление последовательности, и в особенности ТЗУС, могут быть использованы для контроля надежности (обнаружения «взлома») КТОЗУО на идейной основе метода Ра-ша. Для этого в тесте может содержаться некоторое количество легких ТЗ, компенсируемое значением порога положительной оценки, а в критерии положительной оценки – условие превышения значения порога правильных ответов на легкие ТЗ. Дополнительное условие критерия положительной оценки может основываться и на том, что соотношение верных и неверных ответов на легкие ТЗ должно значительно превышать аналогичное соотношение для остальных, более трудных ТЗ или всего КТОЗУО в целом:

$$\frac{N_{\text{легк}}^{\text{отв}}}{N_{\text{легк}}^{\text{вопр}} - N_{\text{легк}}^{\text{отв}}} \geq k_1 \cdot \frac{N_{\text{легк}}^{\text{отв}} + N_{\text{ср}}^{\text{отв}} + N_{\text{тр}}^{\text{отв}}}{N_{\text{легк}}^{\text{вопр}} + N_{\text{ср}}^{\text{вопр}} + N_{\text{тр}}^{\text{вопр}} - (N_{\text{легк}}^{\text{отв}} + N_{\text{ср}}^{\text{отв}} + N_{\text{тр}}^{\text{отв}})} \equiv k_1 \cdot \frac{N^{\text{отв}}}{N^{\text{вопр}} - N^{\text{отв}}}; \quad (2.26)$$

$$\frac{N_{\text{легк}}^{\text{отв}}}{N_{\text{легк}}^{\text{вопр}} - N_{\text{легк}}^{\text{отв}}} \geq k_2 \cdot \frac{N_{\text{ср}}^{\text{отв}} + N_{\text{тр}}^{\text{отв}}}{N_{\text{легк}}^{\text{вопр}} + N_{\text{ср}}^{\text{вопр}} - (N_{\text{ср}}^{\text{отв}} + N_{\text{тр}}^{\text{отв}})} \equiv k_2 \cdot \frac{N^{\text{отв}} - N_{\text{легк}}^{\text{отв}}}{N^{\text{вопр}} - N_{\text{легк}}^{\text{вопр}} - (N^{\text{отв}} - N_{\text{легк}}^{\text{отв}})}. \quad (2.27)$$

Здесь $N^{\text{вопр}}$ и $N^{\text{отв}}$ – общее количество легких вопросов и ответов в КТОЗУО. Коэффициенты k_1 и k_2 в выражениях (2.26) и (2.27) могут быть определены эмпирическим или экспертным путем.

4. Тестовые задания с усложненной структурой требуют большего времени на решение по сравнению с ТЗВО или ТЗАО, поэтому время решения ТОЗУО, содержащих ТЗ с усложненной структурой, уменьшается непропорционально уменьшению количества ТЗ.

5. Тестовые задания с усложненной структурой допускают частично верные ответы: установление части верных соответствий в ТЗУС, правильное указание отдельных позиций элементов последовательности в ТЗУП, выбор одного из правильных вариантов ответа в ТЗМВ, что дает возможность строить сложные оценочные алгоритмы

и шкалы для получения детализированных результатов тестирования индивидуальных и групповых учебных достижений обучающихся применительно к различным предметным областям, ДЕСО и компетенциям на основе сравнительно компактных по объему КТОЗУО.

2.4. Сравнительный анализ надежности тестовых заданий методом экспертного моделирования решения

Для расширения верификационной базы исследований надежности тестов, составленных из однородных ТЗ, методов априорной оценки надежности и достоверности оценок КТОЗУО, нами было проведено экспертное моделирование решения ТЗ различных видов по нескольким предметным областям [87].

В состав экспертной группы по каждой предметной области входило пять преподавателей-предметников, один методист и один психолог. Задачей методиста и психолога, которые работали во всех экспертных группах, было обеспечение единых алгоритмов выработки экспертных оценок и корректности их обсуждения и принятия. Процедура принятия экспертного решения включала в себя выдвижение рабочей гипотезы, ее обсуждение, выдвижение альтернативной гипотезы, обсуждение альтернативной гипотезы, при необходимости корректировку рабочей или альтернативной гипотезы, утверждение одной из гипотез в качестве экспертного решения. Экспертное решение утверждалось исключительно преподавателями-предметниками простым голосованием. Подавляющее большинство (около 90%) решений было принято единогласно, в остальных случаях – при одном воздержавшемся или одном «против», что явилось следствием достижения консенсуса при обсуждении гипотез и их корректировке.

В начале экспертизы преподавателям-предметникам предлагали шестиуровневую шкалу – модель подготовленности тестируемых со следующими характеристиками уровней обученности:

А: тестируемый не владеет учебным материалом, в ответе использует остаточные знания, в том числе из смежных предметных областей, общую эрудицию;

Б: тестируемый имеет лишь общее представление о предметной области, оперирует обрывочными, бессистемными сведениями по изучаемому предмету, использует общую эрудицию;

В: тестируемый обладает фрагментарными знаниями в предметной области в объеме примерно половины дидактических единиц изученного содержания;

Г: тестируемый знает бóльшую часть учебного материала, уверенно решает типовые качественные задачи (простые и среднего уровня сложности);

Д: тестируемый знает весь учебный материал, однако испытывает затруднения в решении сложных качественных задач, в том числе ориентированных на междисциплинарное содержание, ответах на вопросы, требующих углубленных знаний или хороших аналитических навыков;

Е: идеальная модель тестируемого, дающего верные ответы на все рассматриваемые ТЗ.

Суть экспертного решения состояла в определении для рассматриваемого ТЗ и уровня обученности тестируемых объема знаний (частично верных ответов), позволяющих максимально повысить вероятность угадывания верного ответа. Например, в ТЗВО эксперты исключали из рассмотрения некоторое количество дистракторов, соответствующее уровню подготовленности тестируемых, в ТЗУС – некоторое количество пар соответствий и т. д.

Виды рассматриваемых ТЗ:

- 1 – на установление парных соответствий;
- 2 – с выбором единственного верного ответа;
- 3 – с множественным выбором верных ответов;
- 4 – на упорядочивание дистракторов;
- 5 – на установление последовательности;
- 6 – с градацией верных ответов.

Общей задачей экспертного моделирования было выявление на основе сравнительного анализа экспертных оценок видов ТЗ, наиболее устойчивых к угадыванию тестируемыми верного ответа при его реальном незнании. Поэтому составители ТЗ ориентировались на примерно одинаковую содержательную трудность заданий по одной предметной области, а для корректности сопоставления видов ТЗ структура каждого вида содержала пять позиций.

Ниже приведены предложенные экспертам ТЗ шести видов по истории России.

1. Установите соответствие:

Город	Год основания
Владимир	863
Москва	903
Псков	1010
Смоленск	1108
Ярославль	1147

2. Дедом Ивана Калиты был:

- а) Александр Невский;
- б) Юрий Долгорукий;
- в) Симеон Гордый;
- г) Владимир Мономах;
- д) Дмитрий Донской.

3. В числе первых Героев Советского Союза, награжденных за спасение челюскинцев, были:

- а) Александр Покрышкин;
- б) Валерий Чкалов;
- в) Анатолий Ляпидевский;
- г) Иван Доронин;
- д) Герман Титов.

4. Расположите на первом месте правильный ответ, а затем ответы в порядке убывания их достоверности. Первый российский военный корабль был построен:

- а) в Кронштадте;
- б) в окрестностях Коломны;
- в) под Воронежем;
- г) в Китеже;
- д) в Лиссабоне.

5. Расположите в хронологическом порядке фамилии руководителей Правительства СССР:

- а) Иосиф Виссарионович Сталин;
- б) Вячеслав Михайлович Молотов;
- в) Георгий Максимилианович Маленков;
- г) Николай Александрович Булганин;
- д) Алексей Иванович Рыков.

6. Расположите ответы в порядке значимости. Целью строительства Санкт-Петербурга было:

- а) перенос столицы Российского государства из Москвы;
- б) создание плацдарма для ведения Северной войны на Балтике;
- в) освоение северо-западных земель России;
- г) создание крупной торгово-экономической и военной базы, максимально приближенной к европейским торговым центрам;
- д) строительство Балтийского флота.

Принятые экспертами гипотезы по каждому заданию приведены в порядке возрастания моделируемого экспертами уровня обученности тестируемого по заданной шкале.

1 – А: Москва – не 903, Смоленск – не 1108.

1 – Б: Москва – 1147, Смоленск – не 1108, Ярославль – не 863, Владимир – не 863.

1 – В: Москва – 1147, Смоленск – 863, Ярославль – не 903, Псков – не 1108.

1 – Г: Москва – 1147, Смоленск – 863, Псков – 903.

1 – Д = 1 – Г.

2 – А: не Дмитрий Донской.

2 – Б: не Дмитрий Донской, не Юрий Долгорукий.

2 – В: не Дмитрий Донской, не Владимир Мономах, не Юрий Долгорукий.

2 – Г = 2 – В.

2 – Д: верный ответ.

3 – А: не Герман Титов.

3 – Б: не Герман Титов, не Александр Покрышкин.

3 – В: Анатолий Ляпидевский + не Герман Титов, не Александр Покрышкин.

3 – Г: Анатолий Ляпидевский + не Герман Титов, не Александр Покрышкин, не Валерий Чкалов.

3 – Д = 3 – Г.

4 – А: не в Лиссабоне (без определения позиции Лиссабона).

4 – Б: не в Китеже и не в Лиссабоне (без определения позиций Китежа и Лиссабона).

4 – В: 4 – Лиссабон, 5 – Китеж.

4 – Г: 4 – Лиссабон, 5 – Китеж, Кронштадт не на 1-м месте.

4 – Д: 3 – Кронштадт; 4 – Лиссабон, 5 – Китеж.

5 – А: не Сталин на 1-й позиции.

5 – Б: не Сталин, не Маленков на 1-й позиции + Маленков был позже Сталина.

5 – В: не Сталин, не Маленков, не Булганин на 1-й позиции + Маленков и Булганин были позже Сталина.

5 – Г: 1 – Рыков, 2–3 – Молотов и Сталин, 4–5 – Маленков и Булганин.

5 – Д: 1 – Рыков, 2 – Молотов, 3 – Сталин, 4–5 – Маленков и Булганин.

- 6 – А: не а) на 1-м месте.
 6 – Б: не а) и не д) на 1-м месте.
 6 – В: д) – 4; а) – 5; б) не на 1-м месте.
 6 – Г: б) – 3; д) – 4; а) – 5.
 6 – Д: верный ответ.

Рассчитанные для каждого ТЗ и уровня обученности тестируемых вероятности угадывания правильного ответа [69] приведены в табл. 2.1.

Таблица 2.1

Вероятности угадывания правильных ответов
на тестовое задание по истории России

Уровни подготовленности	Типы заданий					
	1	2	3	4	5*	6
А	0,0128	0,25	0,0667	0,0104	0,0104	0,0104
Б	0,1	0,333	0,143	0,0139	0,0294	0,0139
В	0,333	0,5	0,25	0,167	0,0625	0,25
Г	0,5	0,5	0,5	0,25	0,25	0,5
Д	0,5	1	0,5	0,5	0,5	1
Е	1	1	1	1	1	1

Здесь и далее символом «*» отмечены задания, обозначенные экспертами как наиболее трудные.

Аналогичным образом вторая группа экспертов выполнила моделирование решения ТЗ (до момента угадывания верного ответа) по географии.

1. Установите соответствие:

Страна	Столица
Эквадор	Бамако
Малави	Хараре
Мали	Бельмопан
Белиз	Лилонегле
Зимбабве	Кито

2. Самая полноводная река в мире:

- а) Енисей;
 б) Амур;
 в) Амазонка;
 г) Миссисипи;
 д) Нил.

3. Страны, с которыми Россия имеет сухопутные границы:

- а) Узбекистан;
- б) Норвегия;
- в) Польша;
- г) Армения;
- д) Турция.

4. Расположите на первом месте правильный ответ, а затем ответы в порядке убывания их достоверности. Столицей Республики Тыва является:

- а) Улан-Удэ;
- б) Улан-Батор;
- в) Габороне;
- д) Кызыл;
- е) Тана.

5. Расположите города по их близости к Москве:

- а) Казань;
- б) Воронеж;
- в) Минск;
- г) Санкт-Петербург;
- д) Вологда.

6. Расположите ответы в порядке значимости. Значение Северного морского пути для России заключается:

- а) в разведывании месторождений полезных ископаемых в Арктике;
- б) в установлении морского сообщения между Северо-Западом и Дальним Востоком;
- в) в развитии инфраструктуры арктического побережья;
- г) в формировании единой транспортной системы областей страны, связанных реками с Северным Ледовитым океаном;
- д) в развитии международного сотрудничества со странами арктического бассейна.

Результаты экспертного моделирования выглядят следующим образом.

1 – А: Зимбабве – не Кито, Эквадор – не Лилонгве (исключение 2 неверных).

1 – Б: Малави – не Кито, Зимбабве – не Кито, Эквадор – не Лилонгве, Эквадор – не Бамако, Эквадор – не Хараре.

1 – В: Эквадор – Кито; Зимбабве – не Бамако, Зимбабве – не Лилонгве, Белиз – не Хараре, Белиз – не Бамако.

1 – Г: Эквадор – Кито; Зимбабве – Хараре; Белиз – не Бамако, Мали – не Бельмопан.

1 – Д: Эквадор – Кито; Зимбабве – Хараре; Мали – Бамако.

2 – А: не Миссисипи.

2 – Б: не Миссисипи, не Нил.

2 – В: не Миссисипи, не Нил, не Амур.

2 – Г: Амазонка – правильный ответ.

2 – Д = 2 – Г.

3 – А: не Турция.

3 – Б: Польша + не Турция.

3 – В: Польша + не Турция, не Узбекистан.

3 – Г: Польша + не Узбекистан, не Турция, не Армения.

3 – Д: верный ответ.

4 – А: на 1-м месте не Тана и не Габороне.

4 – Б: на 1-м месте не Тана, не Габороне и не Улан-Батор.

4 – В: на 1-м месте Кызыл, на 2-м не Габороне и не Тана.

4 – Г: 1 – Кызыл, 5 – Тана, 2 – не Габороне.

4 – Д: 1 – Кызыл, 4 – Габороне, 5 – Тана.

5 – А: на 1-м месте не Казань и не Воронеж.

5 – Б: на 1-м месте не Казань и не Воронеж, на последнем месте не Санкт-Петербург.

5 – В: на 5-м месте Казань, на 1-м месте не Воронеж и не Минск, Вологда ближе Санкт-Петербурга.

5 – Г: 1 – Вологда, 5 – Казань, Санкт-Петербург – не 4.

5 – Д: 1 – Вологда, 4 – Минск, 5 – Казань.

6 – А: 5-е место – д).

6 – Б: 5-е место – д), 4-е место – а).

6 – В: 5-е место – д), 4-е место – а); г) – не 1-е место.

6 – Г: 1 – б), 4 – а), 5 – д.

6 – Д: верный ответ.

В табл. 2.2 приведены расчетные значения вероятностей угадывания тестируемыми верного ответа для каждого экспертного решения.

Таблица 2.2

*Вероятности угадывания правильных ответов
на тестовое задание по географии*

Уровни подготовленности	Типы заданий					
	1*	2	3	4	5*	6
А	0,0128	0,25	0,0667	0,0139	0,0139	0,0417
Б	0,0278	0,333	0,125	0,0208	0,0417	0,167
В	0,167	0,5	0,25	0,0833	0,167	0,25
Г	0,25	1	0,5	0,25	0,25	0,5
Д	0,5	1	1	0,5	0,5	1
Е	1	1	1	1	1	1

Третьей группе экспертов было предложено провести моделирование частичного или полного решения ТЗ по физике (раздел электричество и магнетизм).

1. Установите соответствие

Физическая величина	Единица измерения
Индуктивность	Тесла
Магнитный поток	Фарад
Проводимость	Генри
Электрическая емкость	Сименс
Магнитная индукция	Вебер

2. Возникновение градиента температур в находящемся в магнитном поле проводнике, через который протекает электрический ток, называется эффектом:

- а) Лебедева;
- б) Пельтье;
- в) Холла;
- г) Эттингсгаузена;
- д) Доплера.

3. Частоту гармонического электрического колебания можно измерить:

- а) осциллографом;
- б) гальванометром;
- в) селективным вольтметром;
- г) частотомером;
- д) мостом Уитстона.

4. Расположите на первом месте правильный ответ, а затем ответы в порядке убывания их достоверности. Относительная ширина полосы частот резонансного контура определяется:

- а) произведением емкости и индуктивности контура;
- б) комплексным сопротивлением контура;
- в) величиной рассеиваемой за период колебаний мощности;
- г) добротностью контура;
- д) надежностью контура.

5. Расположите виды радиочастот в порядке возрастания длины волны:

- 1) КВЧ;
- 2) СВЧ;
- 3) УВЧ;
- 4) ГВЧ;
- 5) ОВЧ.

6. Расположите ответы по их соответствию правильному определению. Ампер в Международной системе единиц (СИ) определяется как:

1) сила тока, который при прохождении по двум параллельным прямолинейным проводникам бесконечной длины и пренебрежимо малой площади кругового поперечного сечения, расположенным в вакууме на расстоянии 1 метр, вызвал бы силу взаимодействия в $2 \cdot 10^{-7}$ ньютона на погонный метр проводника;

2) сила тока, при прохождении которого через стандартный раствор нитрата серебра осаждается 1,11800 мг серебра в секунду;

3) сила тока, при котором через поперечное сечение проводника за одну проходит заряд величиной в 1 кулон;

4) сила тока, который создает падение напряжения в 1 вольт на сопротивлении в 1 Ом;

5) сила тока, который возникает в проводнике сопротивлением 1 Ом при подключении его в качестве нагрузки к источнику ЭДС величиной 1 вольт.

Соответствующие результаты моделирования, закрепленные экспертными решениями, представлены следующими гипотезами о знаниях тестируемых.

1 – А: проводимость – не Фарад, не Генри; электрическая емкость – не Сименс.

1 – Б: электрическая емкость – Фарад; проводимость – не Тесла, не Генри.

1 – В: электрическая емкость – Фарад; индуктивность – Генри; проводимость – не Вебер.

1 – Г: емкость – фарада; индуктивность – Генри; проводимость – Сименс.

1 – Д: верный ответ.

2 – А: не Доплера.

2 – Б: не Доплера, не Холла.

2 – В: не Доплера, не Холла, не Пельтье.

2 – Г = 2 – В.

2 – Д: верный ответ

3 – А: частотомером.

3 – Б: частотомером и осциллографом.

3 – В: частотомером и осциллографом; не гальванометром.

3 – Г: частотомером и осциллографом; не гальванометром, не мостом Уитстона.

3 – Д: верный ответ.

4 – А: 5 – надежностью.

4 – Б: 5 – надежностью, 1 – не комплексным сопротивлением и не произведением емкости на индуктивность.

4 – В: 1 – добротностью, 5 – надежностью, 2 – не рассеиваемой за период колебаний мощностью.

4 – Г: 1 – добротностью, 2 – комплексным сопротивлением контура, 5 – надежностью.

4 – Д = 4 – Г.

5 – А: 5 – не ГВЧ и не СВЧ.

5 – Б: 1 – не СВЧ и не УВЧ, 5 – не СВЧ, не ГВЧ, не КВЧ.

5 – В: 1 – не СВЧ, не УВЧ, не ОВЧ; 2 – не ОВЧ, не УВЧ; 3 – не ГВЧ, не КВЧ.

4 – не ГВЧ, не КВЧ; 5 – не СВЧ, не ГВЧ, не КВЧ.

5 – Г: 1 – ГВЧ или КВЧ, 2 – ГВЧ или КВЧ, 3 – СВЧ или ОВЧ, 4 – СВЧ, УВЧ или ОВЧ, 5 – ОВЧ или УВЧ.

5 – Д: 1 – ГВЧ или КВЧ, 2 – ГВЧ или КВЧ, 3 – СВЧ, 4 – ОВЧ или УВЧ, 5 – ОВЧ или УВЧ.

6 – А: нет вариантов.

6 – Б: 4 – 4) или 5); 5 – 4) или 5).

6 – В = 6 – Б.

6 – Г: 1 – 1), 2 – 2) или 3), 3 – 2) или 3), 4 – 4) или 5), 5 – 4) или 5).
 6 – Д: 1 – 1), 2 – 2) или 3), 3 – 2) или 3), 4 – 4), 5 – 5).

Соответствующие расчетные значения вероятностей угадывания тестируемыми верного ответа приведены в табл. 2.3.

Таблица 2.3

Вероятности угадывания правильных ответов на тестовое задание по физике (электричество и магнетизм)

Уровни подготовленности	Типы заданий					
	1	2	3	4	5*	6*
А	0,0167	0,25	0,0625	0,0417	0,0139	0,00833
Б	0,0833	0,333	0,125	0,0833	0,0333	0,0833
В	0,25	0,5	0,25	0,25	0,05	0,0833
Г	0,5	0,5	0,5	0,5	0,125	0,25
Д	1	1	1	0,5	0,25	0,25
Е	1	1	1	1	1	1

Сравнение данных, представленных в табл. 2.1–2.3, показывает, что уровень обученности «Д» практически гарантирует верное решение, если задание сформулировано в виде вопроса с выбором единственного верного ответа. В одном из трех случаев (география) верный ответ, по мнению экспертов, для данного вида ТЗ обеспечивался уровнем «Г».

При составлении ТЗ наибольшие трудности вызвали ТЗУП, так как необходимо было выдержать единую структуру ТЗ для всех предметных областей и примерно одинаковый уровень содержательной трудности для каждой предметной области. В связи с этим составителям заданий пришлось отказаться от наиболее распространенной структуры ТЗ данного вида – восстановления правильной последовательности слов в определении (задание проигрывало в содержательной трудности остальным ТЗ в предметной области) или символов в формулах (нереализуемо для географии и истории России). В итоге ТЗУП, единые по структуре для всех трех предметных областей, обладали некоторым сходством с ТЗУС. Однако, по мнению экспертов, алгоритмы решения ТЗУП и ТЗУС существенно отличались: позиции в последовательностях формировались путем взаимного сопоставления относительно общего признака, что не требовало конкретики знаний как отправной точки решения, в отличие от ТЗУС. Данный под-

ход к составлению ТЗУП привел к повышению их содержательной трудности во всех трех предметных областях.

Несмотря на достаточное количество дистракторов (четыре), ТЗВО оказались подвержены угадыванию верного ответа обучающимися, чьи модели обученности характеризуются неполнотой и даже фрагментарностью усвоения тестируемого содержания.

Усложнение ТЗМВ за счет заранее неизвестного тестируемым количества верных ответов не привело к существенному снижению вероятности угадывания верных ответов.

По мнению экспертов, ТЗУД и ТЗУП ни по одной из предметных областей не могли быть корректно решены обучающимися, чей уровень подготовки позволял бы им получать преимущественно отличные оценки по соответствующим учебным дисциплинам. Данный результат объяснен экспертами исключительно специфической структурой данных ТЗ, требующих от обучающихся одновременно конкретных знаний и понимания их содержательной и логической взаимосвязи.

Обеспечение трудности ТЗВО, ТЗМВ и ТЗУП, необходимой трудности ТЗУД и ТЗУС оказалось возможным лишь за счет повышения содержательной трудности заданий. Трудность ТЗГО и ТЗУД оказалась зависящей от предметной области и наивысшей для естественно-научных дисциплин.

По мнению разработчиков тестов и экспертов, имеет место корреляция между сложностью разработки ТЗ и трудностью их решения, что особенно характерно для ТЗУП и ТЗГО.

3. НЕКОТОРЫЕ ВОПРОСЫ ПОВЫШЕНИЯ НАДЕЖНОСТИ И ДОСТОВЕРНОСТИ ТЕСТИРОВАНИЯ УЧЕБНЫХ ДОСТИЖЕНИЙ

3.1. Реализация целеполагания контроля учебных достижений обучающихся в разработке дидактических тестов

При КУДО в форме собеседования квалификация преподавателя – его компетентность в предметной области и педагогическое мастерство – наряду с высокими моральными качествами играют определяющую роль не только в достижении высокого уровня достоверности оценивания учебных достижений обучающихся, но и в их мотивации к успешному освоению содержания обучения, контролируемого как в текущий момент, так и в перспективе. ТОЗУО, аккумулируя знания, опыт и педагогические компетенции разработчика тестов или целого коллектива педагогов, переносит их в множественные ситуации КУДО, тем самым компенсируя возможный недостаток дидактических компетенций субъекта контроля.

Достоверность оценок и надежность КТОЗУО обеспечиваются достоверностью и надежностью оценок ТЗ в отдельности, корректностью алгоритмов оценивания, включая формирование выборки ТЗ из их базы (структура выборки, распределение ТЗ по уровням трудности, разнообразие видов ТЗ), начисление баллов за полностью и частично правильно решенные задания с учетом их трудности, применение штрафов, выбор порогов и диапазонов оценок, перевод набранных баллов в оценки (в целом теста и отдельных компетенций, тем, дидактических единиц содержания) и др., а также корректностью моделей и процедур обобщения и первичного анализа данных.

Проблемы достоверности КТОЗУО могут быть сведены к ошибкам оценивания выполнения ТЗ первого рода (расценивание ответа как неверного при реальном знании верного ответа испытуемым) и ошибкам второго рода (засчитывание ответа как верного при его незнании) [26, 77]. Причинами ошибок первого рода являются методические и технические ошибки разработчика теста, невнимательность испытуемых при прочтении заданий и вводе ответов и их низкая мотивация к решению тестов, например при контроле остаточных знаний в ходе контрольных мероприятий, выходящих за рамки учебного плана (внешнего аудита образовательной организации, мониторинга реализации образовательных программ в вузовских системах управления качеством образования и др.). Характерными методическими

недостатками тестов, приводящими к ошибкам первого рода, являются [30, 35, 73, 82, 83]:

избыточное видовое многообразие ТЗ наряду с частым чередованием ТЗ различных видов, требующее постоянного переключения внимания испытуемых на изменяющийся порядок их выполнения;

незначительность или малозаметность отличий дистракторов от верного ответа (замена термина равнозначным эквивалентом, относительно малые количественные несовпадения и т. п. (примеры 3.1, 3.2);

избыточное количество дистракторов в ТЗ с выбором единственного верного ответа или нескольких верных ответов (пример 3.1);

громоздкие формулировки вопросов и вариантов ответов, требующие дополнительных усилий и времени на их внимательное прочтение и осмысление;

избыточная трудность ТЗ, несоизмеримая со временем их выполнения;

относительно большое количество ТЗ, направленных на проверку малозначительных знаний, частных, второстепенных деталей (пример 3.2);

некорректное начисление баллов и штрафов при частично верном решении в ТЗМВ, ТЗУС, ТЗГО, ТЗУП, ТЗУД;

неоднозначность выбора верного ответа из предлагаемых вариантов, причиной которой могут быть частично верные дистракторы, спорность верного ответа (пример 3.1);

опора на субъективные суждения или субъективное восприятие каких-либо событий и явлений (пример 3.3).

Пример 3.1. Первая Гагская конференция завершилась:

- 1) 11 июня 1899 г.;
- 2) 23 июня 1899 г.;
- 3) 5 июля 1899 г.;
- 4) 17 июля 1899 г.;
- 5) 29 июля 1899 г.;
- 6) 10 августа 1899 г.

Пример 3.2. Применение химических снарядов запрещено:

- 1) конвенцией, принятой Первой Гагской конференцией;
- 2) декларацией, принятой Первой Гагской конференцией;
- 3) манифестом, принятым Первой Гагской конференцией;
- 4) хартией Первой Гагской конференции;
- 5) итоговым протоколом Первой Гагской конференции.

Пример 3.3. Какое сражение Наполеон назвал самым страшным из всех, в которых он принимал участие?

- 1) под Бородином;
- 2) под Лейпцигом;
- 3) под Аустерлицем;
- 4) под Берлином [83].

Ошибки второго рода, приводящие к завышению оценок реальных знаний обучающихся, чаще всего проявляются при контроле текущей успеваемости, предварительной и промежуточной аттестации. Их основные причины – знание испытуемыми части верных ответов заранее (например, при многократном применении теста, когда информация о вопросах и формулировках предположительно верных ответов может передаваться от прошедших тестирование обучающихся к тем, чьи знания еще не подвергались проверке) и недостаточная устойчивость ТЗ к угадыванию верных ответов, порожденная, в свою очередь, методическими просчетами составителя теста. Наиболее типичные из них [8, 26, 35, 82]:

недостаточное количество дистракторов в ТЗВО, что повышает вероятность случайного угадывания (пример 3.4);

скрытые подсказки в виде явных отличий (фразеологических, синтаксических, лексических) верных ответов от дистракторов (пример 3.5);

скрытые подсказки в виде абсурдности отдельных дистракторов или тривиальности их распознавания и исключения из рассмотрения испытуемым (пример 3.6);

скрытые подсказки в формулировках вопросов (пример 3.7);

взаимосвязь содержания ТЗ, когда какой-либо вопрос является ответом (или наводит на ответ) на другой вопрос (пример 3.8);

относительно большое количество тривиальных вопросов и заданий, а также вопросов, ответы на которые не требуют понимания изученного материала или могут быть даны на основе общей эрудиции (пример 3.4).

Пример 3.4. Какие угрозы безопасности информации являются преднамеренными?

- 1) несанкционированный доступ;
- 2) ошибки персонала;
- 3) открытие электронного письма, содержащего вредоносную программу.

Пример 3.5. Обоснованность отнесения сведений к государственной тайне и их засекречивание заключается:

1) в установлении путем экспертной оценки целесообразности засекречивания конкретных сведений, вероятных экономических и иных последствий этого акта, исходя из баланса жизненно важных интересов государства, общества и граждан;

2) в решении руководителя органа внутренних дел;

3) в проверке соответствия сведений пункту перечня сведений, подлежащих засекречиванию;

4) в проставлении в установленном порядке соответствующего грифа секретности.

Пример 3.6. Командир монгольского отряда, ответственный за сбор дани на Руси и контролирующий князей, назывался:

1) баскак;

2) пристав;

3) ярлык;

4) туарег;

5) виконт.

Пример 3.7. Поражение под Нарвой, победа под Полтавой, победа в Гангутском морском сражении относятся к событиям:

1) русско-польской войны 1654–1667 гг.;

2) Северной войны;

3) русско-турецкой войны 1768–1774 гг.;

4) Семилетней войны.

Пример 3.8. Вопрос 1. Основное содержание деклараций Первой Гаагской конференции:

1) запрещение применения отдельных видов боеприпасов;

2) пересмотр Женевских конвенций 1864 г.;

3) законы и обычаи сухопутных войн;

4) права и обязанности нейтральных государств;

5) статус и права военнопленных.

Вопрос 2. Боевое применение пуль со смещенным центром тяжести запрещено декларацией Первой Гаагской конференции:

1) накануне англо-бурской войны;

2) после русско-японской войны;

3) по окончании Первой мировой войны;

4) накануне Балканских войн;

5) по окончании Крымской войны.

Требования к разработке тестов и ТЗ, направленные на минимизацию ошибок оценивания первого и второго рода, довольно противоречивы. Так, минимизация ошибок второго рода за счет обеспечения приемлемой дифференцирующей способности теста требует

увеличения количества ТЗ и дифференциации их трудности, что является фактором снижения надежности ТОЗУО, приводит к увеличению времени его выполнения и тем самым к повышению вероятности ошибок первого рода.

Снижению ошибок первого рода способствует видовая однородность ТЗ, однако составление ТОЗУО из ТЗ одного вида создает трудности в обеспечении его дифференцирующей способности при ограниченном количестве заданий.

Факторами обеспечения надежности КТОЗУО являются воспроизводимость результатов тестирования и устойчивость теста к деградации. Однако воспроизводимость результатов тестирования снижается с повышением вариативности выборки ТЗ, т. е. увеличением и редактированием их базы, которые, в свою очередь, обеспечивают уменьшение вероятности ошибок второго рода за счет деградации КТОЗУО.

Повышение порога положительной оценки, с одной стороны, приводит к снижению вероятности ошибок второго рода, с другой стороны, за счет сужения диапазонов положительных оценок – к уменьшению дифференцирующей способности теста, также обеспечивающей минимизацию ошибок второго рода.

Таким образом, разработчик теста, оперируя механизмами достижения искомых показателей иногда противоречивых характеристик тестов, по сути, решает задачу многомерной оптимизации. Оптимизируемыми параметрами являются количество ТЗ, их трудность, время тестирования, количество начисляемых баллов за ответы (в том числе частично верные) на задания различной трудности и величина штрафов за неверные ответы, диапазоны правил перевода баллов в оценки. Критерии оптимальности и оптимизации, их реализация в алгоритмах тестирования, как правило, носят эмпирико-эвристический характер, а достижение приемлемой достоверности оценок тестирования определяется квалификацией разработчика теста, зависящей от его практического опыта и умения оперативно и творчески реагировать на возникающие проблемы применения КТОЗУО, связанные, например, с доступностью для испытуемых чатов ИИ и технологий скрытного приема и передачи информации.

Как показывает экспертиза тестов и баз ТЗ в ФОС, основной и первопричиной недостоверного оценивания учебных достижений обучающихся является неопределенность целеполагания при разработке тестов, что довольно характерно и для формирования ФОС в целом. Неопределенность целеполагания может проявляться:

в неопределенности целевой аудитории (отсутствии «привязки» содержания и трудности КТОЗУО по учебной дисциплине к специальности и специализации тестируемых, учета содержания ранее изученных ими учебных дисциплин);

в отсутствии или неправильном учете приоритетов наиболее важных ДЕСО в объеме ТЗ;

в несоответствии содержания ТЗ и их трудности виду контроля учебных достижений: предварительной или промежуточной аттестации, контролю остаточных знаний (с учетом сроков завершения изучения контролируемого содержания), оцениванию уровня сформированности компетенции и др.;

в отсутствии априорного моделирования тестирования учебных достижений обучающихся, основанного на модели обученности тестируемых, необходимого для корректировки дифференцирующей способности теста и оценки его надежности.

Иными словами, разработчики тестов для КУДО зачастую создают их универсальными, применимыми в любой ситуации, ориентируясь главным образом на охват контролируемого содержания безотносительно целевой аудитории и особенностей целеполагания, способных через трудность теста, его дифференцирующую способность и другие характеристики существенно повлиять на выбор видов ТЗ, распределение их трудности, объем и вариативность выборки и иные компоненты структуры теста.

Отправной точкой конкретизации задач разработки ТОЗУО может стать модель испытуемого. Разработчику теста следует не только ориентироваться на предполагаемый уровень обученности испытуемых (в том числе с учетом времени, прошедшего с момента окончания изучения контролируемого содержания), но и количественно определить для себя модельные параметры той части контингента обучающихся, которая должна остаться за чертой удовлетворительной оценки.

Модель аутсайдера должна учитывать не только приемлемый уровень знаний, выраженный в количестве набранных баллов или проценте правильных ответов, но и тактику выполнения ТЗ с неизвестным ответом и наиболее вероятный сценарий решения теста в целом. Примерная классификация учебных достижений и когнитивных способностей обучающихся, которая с учетом перспектив ее детализации может быть положена в основу построения моделей испытуемых, включает в себя следующие категории:

1) «знает, соображает» – обучающийся не только освоил содержание обучения, но и умеет применить его для решения задач, в том числе комплексного характера;

2) «знает, не соображает» – обучающийся освоил содержание обучения формально («вызубрил»), затрудняется с применением знаний, особенно при видоизменении заданий, допускает ошибки в интерпретации полученных решений, не в полной мере ориентируется в причинно-следственных и логико-структурных взаимосвязях отдельных ДЕСО;

3) «не знает, соображает» – обучающийся не освоил содержание обучения или его большую часть, однако за счет общих знаний и эрудиции, профессионального кругозора, знаний смежных дисциплин, жизненного опыта и интеллекта способен истолковать отдельные термины и понятия, сориентироваться в отдельных темах и ДЕСО, в особенности с помощью наводящих вопросов и подсказок;

4) «не знает, не соображает» – обучающийся не освоил содержание обучения или его большую часть, включая тезаурус, и при этом не обладает достаточным интеллектом и знаниями в смежных областях, чтобы применять их в совокупности с фрагментарными, обрывочными, бессистемными знаниями контролируемого материала для успешного прохождения соответствующих аттестационных испытаний.

Очевидно, проектирование КТОЗУО должно быть нацелено на положительное оценивание знаний испытуемых первой из названных категорий и оценивание как неудовлетворительных знаний испытуемых четвертой категории, а также сведение к минимуму вероятности получения положительной оценки испытуемыми второй и третьей категорий.

Для повышения объективности оценивания учебных достижений испытуемых второй категории ТЗ, помимо проверки знаний, должны выявлять их глубину и системность. Здесь инструментальным подспорьем разработчика КТОЗУО могут стать ТЗ усложненной структуры: ТЗГО, ТЗУД [8, 26, 73, 84].

Объективность оценивания учебных достижений испытуемых третьей категории должна основываться на конкретике проверяемых знаний, основу соответствующего теста могут составлять относительно простые разновидности ТЗ: ТЗВО и ТЗМВ. Напротив, ТЗУС и ТЗУП как объект эффективного применения фрагментарных знаний и общей эрудиции вместе с логикой испытуемых данной категории желательно исключить из КТОЗУО.

Разные авторы [1, 6, 20, 34, 35, 59], исходя из целеполагания разработки оценочных средств, помимо диагностических тестов, под-

разделяющихся на тесты общих и специальных способностей и тесты обученности и учебных достижений, выделяют в отдельный класс дидактические тесты с особым отбором содержания ТЗ, предназначенные не только для выявления знания или незнания изученного материала, но и для оценивания прочности знаний, их полноты, глубины, системности, актуальности, конкретности и общности. К особенностям разработки дидактических тестов, по нашему мнению, следует отнести:

системность выборки ТЗ, которая обеспечивается как системным характером формирования базы заданий, так и алгоритмом формирования выборки;

детализацию оценивания наиболее важных, ключевых ДЕСО;

неявное (в том числе с использованием ТЗ различных видов и различной трудности, с формулировками «тезис – антитезис» и др.) дублирование проверяемых ДЕСО в одной тестовой выборке;

применение ТЗ, ориентированных на проверку понимания, а не формального знания материала;

применение различного рода терминологических и смысловых «ловушек», способных выявить «зазубривание» материала без глубокого понимания его смысла;

применение проверочных «закладок», способных выявить угадывание правильных ответов при их реальном незнании или деградацию базы ТЗ [26, 29];

штрафы за попытки угадать верный ответ в алгоритмах начисления баллов за выполнение ТЗ (например, вычитание баллов за неверный ответ и неначисление баллов за отказ от выбора ответа (пропуск задания));

начисление большего количества баллов за верное решение ТЗ, требующих системного понимания изученного материала.

Воспроизведение всех перечисленных выше требований в одном тесте приведет к увеличению его объема (за счет дублирования проверки одних и тех же ДЕСО, увеличения количества ТЗ для проверки наиболее важных ДЕСО и диагностики деградации теста) или, в случае применения ТЗ усложненной структуры, позволяющих повысить дифференцирующую способность теста без увеличения его объема, к усложнению алгоритма оценивания и разработки теста в целом. Проектирование диагностических тестов требует высокой квалификации разработчика, опыта применения им ТЗ различных видов, владения методиками априорного оценивания надежности теста, уровня его трудности и дифференцирующей способности, а также статистиче-

ского анализа эмпирических данных для корректировки и доработки теста.

Несмотря на высокий диагностический и дидактический потенциал ТЗУС, ТЗУП, ТЗГО, ТЗУД, подавляющее большинство разработчиков КТОЗУО в качестве основной структурной единицы теста отдают предпочтение ТЗВО [30]. В связи с этим сформулируем требования к тестам из ТЗВО, которые призваны придать диагностическому тесту дидактическую направленность:

однородность (стилистическая, терминологическая, смысловая) предлагаемых вариантов ответа каждого ТЗ в отдельности [8, 30, 52] (пример 3.9);

применение ТЗ с удвоенным противопоставлением [1, 52, 59] (пример 3.10);

применение в ТЗ сочетаний одинаковых или близких по смыслу слов, знаков, терминов [83] (пример 3.11);

формулирование ТЗ на основе импликации, т. е. проверки знаний причинно-следственных отношений объективного характера с использованием в явном или неявном виде логической связки «если... то...» [52, 59, 74] (пример 3.12);

использование в части ТЗ вариантов ответа «все ответы верные» или «все ответы неверные» в качестве как дистрактора, так и верного ответа;

комплексный характер формулирования ТЗ, обеспечивающий проверку взаимосвязанных знаний, понятий [83] (вопрос 2 примера 3.8);

применение в качестве дистракторов слов, созвучных верному ответу [83] (пример 3.6а);

сочетание ТЗ с короткими, а именно на знание, узнавание (пример 3.6а), и развернутыми, в частности на знание, понимание, рассуждение (пример 3.4а), дистракторами [83].

Пример 3.9. Аттестация объектов информатизации по требованиям безопасности не проводится в отношении:

1) автоматизированных систем, средств связи, обработки и передачи информации;

2) помещений, предназначенных для ведения секретных и конфиденциальных переговоров;

3) технических средств, установленных в выделенных помещениях;

4) физических лиц, имеющих доступ к объекту информатизации;

5) стационарных устройств хранения информации.

Пример 3.10. В битве на Куликовом поле в 1380 г. сражались войска:

- 1) Чингисхана и Тамерлана;
- 2) Тохтамыша и Дмитрия Донского;
- 3) Мамай и Тохтамыша;
- 4) Александра Невского и Батые;
- 5) Дмитрия Донского и Мамай.

Пример 3.11. К способам защиты информации относятся:

- 1) правовой, криптографический, технический, физический;
- 2) организационный, правовой, технический, тактический;
- 3) физический, технический, аналитический, программный;
- 4) индивидуальный, объектный, коллективный, всеобщий;
- 5) программно-технический, организационно-правовой, структурно-логический.

Пример 3.12. Снижение процентной ставки по кредитам приводит:

- 1) к росту инфляции;
- 2) к активизации предпринимательской деятельности;
- 3) к увеличению объемов банковских вкладов;
- 4) к снижению потребительского спроса;
- 5) к притоку в экономику банковского капитала.

Пример 3.4а. Какие угрозы безопасности информации являются преднамеренными?

- 1) несанкционированный доступ;
- 2) халатное отношение администратора системы информационной безопасности к своим обязанностям;
- 3) открытие электронного письма, содержащего вредоносную программу;
- 4) разглашение паролей и кодов доступа третьим лицам;
- 5) работа с несертифицированным оборудованием и программным обеспечением.

Пример 3.6а. Командир монгольского отряда, ответственный за сбор дани на Руси и контролирующий князей, назывался:

- 1) баскак;
- 2) ясак;
- 3) басурман;
- 4) басмач;
- 5) темник.

Конкретизация и детализация целей КУДО неизбежно приводит к увеличению объемов тестов, в особенности с учетом различных

моделей решения отдельных ТЗ и КТОЗУО в целом, характерных для разных категорий испытуемых. В связи с этим разработчик КТОЗУО, опираясь на собственный опыт преподавания и знание контингента испытуемых, может ориентировать КУДО на минимизацию ошибок оценивания второго рода наиболее многочисленной категории испытуемых.

Направленность КТОЗУО на определенную модель обучающегося, реализованная в недостижимости порога удовлетворительной оценки для соответствующей целевой категории или группы испытуемых, призвана не столько предъявить обучающимся объективные результаты освоения ими образовательных программ, сколько продемонстрировать и конкретизировать им проблемы в обучении, порождающие пробелы в знаниях и умениях. Тем самым результаты КУДО с применением дидактических тестов и их анализ могут стать отправной точкой для осознанного выбора обучающимися наиболее эффективных дидактических средств образовательной организации, перехода на интерактивные модели обучения с возрастанием ролевого статуса обучающихся в образовательном процессе.

3.2. Применение тестовых заданий с повышенной устойчивостью к возникновению ошибок второго рода

Проблемой разработки и применения КТОЗУО в современных условиях следует считать их уязвимость к использованию испытуемыми в неблагоприятных целях современных информационно-коммуникационных технологий и ТИИ, что обуславливает необходимость применения дополнительных мер по обеспечению достоверности оценок знаний испытуемых на контрольных и конкурсных мероприятиях в условиях новых информационно-технологических возможностей влияния на результаты оценивания.

Одним из путей решения данной проблемы является повышение интеллектуального уровня КТОЗУО при одновременном сохранении объема проверяемого содержания. Недостатком ТЗВО является то, что верный ответ испытуемого не гарантирует понимания им проверяемого содержания, причинно-следственных связей и другого и может быть дан в результате заучивания или угадывания за счет сокращения до минимума возможных вариантов ответа на основе иных знаний [29, 87]. Следовательно, ТЗ, сохраняя оперативность их выполнения и проверки, должны обеспечивать достоверность оценивания

не только знания, но и понимания содержания испытуемым, умения им анализировать и синтезировать знание в предметной области.

Очевидным решением обеспечения высокой достоверности оценок КТОЗУО является увеличение количества ТЗ в выборке, что создает максимальный охват контролируемого содержания обучения и возможность варьирования пороговыми значениями оценочных критериев в широких пределах. Однако на практике разработчику КТОЗУО приходится решать задачу оптимизации объема выборки ТЗ на основе поиска компромисса между полнотой оцениваемых ДЕСО и временем тестирования, что опосредованно определяет надежность оценок знаний и умений обучающихся. Поиск такого компромисса основан на разрешении противоречия между необходимостью увеличения количества ТЗ и соблюдением некоторого оптимального времени тестирования, превышение которого ведет к росту вероятности ошибок первого рода [26]. Данное противоречие обуславливают восстребованность расширенного применения в КТОЗУО ТЗ с усложненной структурой: ТЗУП, ТЗГО, ТЗУД, которые характеризуются повышенной устойчивостью к деградации и ошибкам второго рода [8, 22, 30, 54]. Таким образом, увеличение количества ТЗ с усложненной структурой является альтернативным и эффективным способом повышения надежности КТОЗУО.

Тестовые задания с упорядочиванием дистракторов более устойчивы к ошибкам второго рода, чем ТЗВО, ТЗМВ, ТЗУС, ТЗУП, т. е. КТОЗУО, составленный из ТЗУД, потенциально более надежен по сравнению с КТОЗУО на основе других видов ТЗ [26, 80, 81, 87]. Дополнительный интерес к анализу вероятности угадывания верных ответов на ТЗУД вызван и высоким дидактическим потенциалом ТЗ данного вида, выражающимся в их высокой дифференцирующей способности, расширенных возможностях детализированного анализа выявленных знаний, дополнительной обучающей функции [26, 56].

Пример 3.13. Какая из перечисленных планет Солнечной системы наиболее удалена от Солнца?

- 1) Уран;
- 2) Оберон;
- 3) Нептун;
- 4) Орион;
- 5) Меркурий.

Правильная последовательность, вводимая тестируемым с клавиатуры: 31524. Объяснение ответа: 3) Нептун – абсолютно и единственно верный ответ; 1) Уран – 7-я планета Солнечной системы,

наиболее удаленная от Солнца, не считая Урана; 5) Меркурий – единственная из оставшихся трех наименований планета Солнечной системы, пусть и ближайшая к Солнцу; 2) Оберон – спутник Урана, не планета Солнечной системы, но ее элемент; 4) Орион – созвездие, астрономический объект, не имеющий отношения к Солнечной системе.

Повышенная устойчивость теста к деградации, проявляющейся в ошибках второго рода, обеспечивается тем, что последовательность вариантов ответа генерируется программой тестирования каждый раз заново, поэтому правильные последовательности будут раз от раза отличаться, что делает бесполезным их запоминание обучающимися. Знание правильного ответа, т. е. номера первой позиции, может, в зависимости от правил начисления баллов за частично верные ответы, не гарантировать положительной оценки за решение ТЗ в целом, а определение расположения ответов по порядку методом проб и ошибок без достаточных знаний предмета потребует большого количества попыток. Если у преподавателя возникнут основания полагать, что ТЗУД за время использования критически деградировало, он может легко восстановить его работоспособность, изменив некоторые варианты ответа. Для примера 3.13 измененная последовательность вариантов ответа может выглядеть следующим образом: 1) Титан; 2) Сатурн; 3) Уран; 4) Плутон; 5) Нептун. Рассуждения, приводящие к верной последовательности (53214), будут заметно отличаться от приведенных выше и потребуют новых знаний об устройстве Солнечной системы.

В ТЗВО и ТЗМВ замена одних неверных ответов на другие практически не снижает вероятность ошибки второго рода в тесте, подвергшемся «взлому», и для повышения надежности единичных ТЗ требуется полная переработка задания. В рассмотренном примере ТЗУД сама усложненная структура ответа обеспечивает относительную долговечность задания за счет регулярного редактирования дистракторов.

В [26] представлен алгоритм расчета итогового балла при решении КТОЗУО, составленного из 10 ТЗУД, содержащих по 5 вариантов ответов на вопрос. Алгоритм включает себя:

- 1) правило начисления баллов и условие успешного решения теста;
- 2) определение вероятности правильного указания позиции ответа для одного ТЗ;

3) расчет среднего значения вероятности $\bar{p}_j$ угадывания верной j-й позиции во всех заданиях КОТЗУО, среднего набираемого балла по каждой позиции и среднего суммарного балла;

4) вычисление вероятности прохождения КТОЗУО;

5) расчет прогнозируемого итогового балла как суммы произведений вероятностей, рассчитанных по формуле Бернулли [69], на соответствующие им баллы.

Модели тестируемых были представлены тринадцатью уровнями подготовки (включая нулевой – угадывание вслепую), которые предполагают разную среднюю вероятность (разным процентом) знания правильных позиций вариантов ответа на вопросы ТЗ, в том числе всех первых (абсолютно верных ответов) и пятых (абсолютно неверных ответов), что указывает на частичную деградацию КТОЗУО.

Результаты моделирования показывают следующее:

тест, составленный из 10 ТЗУД, обеспечивает меньшее количество набираемых за счет угадывания баллов (в среднем не более 10% от максимального балла) и меньшую вероятность получения удовлетворительной оценки тестируемыми с расчетным гарантированным их уровнем подготовки баллом, меньшим порогового значения удовлетворительной оценки, по сравнению с тестами из 20 ТЗ любых других типов [29, 80, 81], что позволяет формировать на основе ТЗУД компактные тесты повышенной надежности;

структура ТЗУД препятствует получению незаслуженной удовлетворительной оценки даже тестируемым с изначально высоким гарантированным баллом, если их реальные знания не обеспечивают решение более 60% заданий теста.

В ходе моделирования были применены два алгоритма начисления баллов (со своим порогом положительной оценки каждый), учитывающие значимость позиции эталонной последовательности, один из которых предполагал начисление только положительных баллов за совпадение соответствующей позиции вводимой испытуемым последовательности с позицией эталона, а второй начислял штрафы за неверное указание номеров позиций вариантов ответа.

Во втором случае даже изначально превышение гарантированным баллом величины порога положительной оценки при начислении штрафов за неверные ответы не всегда обеспечивает итоговую положительную оценку.

По сравнению с тестируемыми со знанием верных ответов, в общем больше шансов на успех имеют тестируемые с более глубокими знаниями учебной дисциплины даже при меньшем тематическом

охвате. Дополнение процедуры расчета итогового балла начислением штрафов за неверные ответы усиливает дифференциацию оценок подготовки тестируемых и сводит к минимуму ошибки второго рода за счет угадывания верных ответов без их реального знания, тем самым позволяя достоверно выявить не только не освоивших учебный материал обучающихся, но и проблемные для освоения ДЕСО.

Алгоритм оценивания ТЗУД с начислением штрафных баллов обладает большей устойчивостью к «взлому» баз знаний и угадыванию верных ответов. Совершенствование данного алгоритма возможно посредством учета при начислении штрафа степени отклонения цифр в ответе тестируемого на ТЗУД от их истинных значений, т. е. корреляции вводимой и эталонной последовательностей. При этом особое внимание следует уделить первой и последней позициям последовательности ответа [56].

Алгоритм с начислением штрафных баллов востребован в рамках промежуточной аттестации в форме экзамена как обеспечивающий объективную дифференциацию положительных оценок или на завершающем этапе тестирования, которому предшествует КТОЗУО без начисления штрафа или на основе ТЗ с более простой структурой, а также с учетом возможности регулировки порога положительной оценки и параметров, характеризующих значимость позиций эталонной последовательности, перспективен для использования в адаптивных тестах [26, 32, 85, 86].

Редкое использование ТЗУД в КУДО в значительной мере обусловлено относительно недавним созданием и пропагандированием данного вида тестовых заданий [26, 30, 55, 56] и, как следствие, отсутствием «конструкторов» ТЗУД в тестовых блоках систем управления обучением (например, СУО «Moodle»). Крайне редко в тестирующих программах предусмотрены и конструкторы ТЗГО. Однако, поскольку структура ответа на ТЗУД или ТЗГО представляет собой последовательность символов (цифр или букв) при его непосредственном вводе или последовательном выборе позиций дистракторов или, соответственно, верных ответов, создание ТЗУД и ТЗГО возможно в конструкторах ТЗ со сходным синтаксисом ответа: ТЗУП и в ряде случаев ТЗУС.

Второй причиной относительно редкого использования ТЗУД и ТЗГО в тестах для КУДО является ограниченный, по сравнению, например, с ТЗМВ или ТЗАО, набор задач и вопросов, на основе которых могут быть сформулированы ТЗУД и ТЗГО. В большей степени это относится к ТЗГО, так как формулирование ответов с различ-

ной степенью близости к истинному требует не только глубокого и всестороннего знания предмета, но и определенного навыка частичной фальсификации верного ответа [1, 30]. Как правило, и то, и другое формируется по мере приобретения педагогического и исследовательского опыта в предметной области. Проблемой создания ТЗГО является объективность оценивания ответов, особенно частично верных. Отнесение варианта ответа к первой позиции (самый верный ответ) обычно не столь неоднозначно, но размещение вариантов на последующих позициях может оказаться небесспорным в любых учебных дисциплинах. Для получения полностью верного ответа на ТЗГО испытуемый должен не только знать проверяемое содержание учебной дисциплины, но и в процессе градации предлагаемых вариантов ответа в своих рассуждениях воспроизвести логику составителя тестового задания, согласовать свои оценки приоритетов вариантов ответа с его трактовками.

Пример 3.14. Основная цель первой экспедиции Колумба (1492–1493 гг.):

- 1) открытие новых торговых путей в Индию;
- 2) присоединение к владениям испанского короля новых территорий;
- 3) усиление политического влияния Испании на Западе Европы;
- 4) удовлетворение амбиций Колумба как мореплавателя;
- 5) личное обогащение Колумба и его сподвижников.

Градация вариантов ответа на данный вопрос основывается на трех подходах к расстановке приоритетов: первый – личные мотивы Колумба, второй – государственные интересы Испании, третий – всеобщее выгодное развитие торговли.

Соответственно, последовательность вариантов ответа в ТЗГВО на основе первого подхода: 5 4 1 2 3. Здесь мы руководствуемся принципом «от частного к общему», т. е. от личных интересов к государственным через групповые, при этом делаем допущение, что стремление к обогащению у Колумба доминировало над желанием прославиться, что далеко не лишено оснований [91].

Реализация второго подхода даст нам обратную последовательность вариантов ответа: 3 2 1 4 5 (с учетом указанного выше допущения).

Наконец, третий подход определяет первую позицию в последовательности варианта ответа 1. Далее решение ветвится. С одной стороны, для снаряжения и отправки экспедиции было необходимо созревание благоприятной политической ситуации, в частности поглощение

Испанией Гранадского халифата, т. е. переориентация государственных интересов на конкуренцию в морской торговле с Португалией. Данное рассуждение приводит нас к последовательности 1 2 3 5 4. Расстановка вариантов ответа 2 и 3 определяется соображениями приоритета достижения прагматичных, реализуемых в обозримой перспективе целей над глобальными целями, достижение которых зависит от множества политических и экономических условий, а также субъективных факторов. С другой стороны, именно беспрецедентная активность Колумба, основанная на личной заинтересованности, сделала данную экспедицию возможной (по крайней мере, в XV в.), а ее относительный успех (возвращение двух кораблей из трех) также обеспечен умелым руководством и относительной осторожностью великого мореплавателя, для которого на кону стояло сохранение титула вице-короля Индии и сопутствующих привилегий, дарованных ему королем Испании. Следствием данного рассуждения является последовательность вариантов ответа: 1 5 4 2 3.

Таким образом, налицо четыре аргументированных решения, каждое из которых, в свою очередь, допускает вариативность отдельных позиций. Во избежание такой ситуации при формулировании ТЗГО следует основываться на единственном подходе или принципе, а варианты частично верных ответов чаще всего характеризуются различной степенью неполноты по отношению к варианту, располагаемому на первой позиции.

Пример 3.15. Относительная величина потерь мощности в последовательном резонансном контуре определяется значениями:

- 1) добротности контура;
- 2) добротности катушки индуктивности;
- 3) реактивного сопротивления контура;
- 4) индуктивности.

Ответы расположены в правильном (определенном разработчиком тестового задания) порядке. Позиции 2 и 3 установлены на основе того, что добротность контура в основном определяется добротностью катушки индуктивности, как правило, намного меньшей добротности конденсатора.

Более строгое следование принципу убывающей полноты соответствия при формулировании вариантов частично верных ответов, как правило, сопряжено с уменьшением вариантов таких ответов, т. е. приводит к увеличению вероятности угадывания верной последовательности и снижению надежности ТЗ и теста в целом. Зачастую неполнота ответа заметна визуально или при поверхностной оценке:

по количеству слов или перечисляемых значимых признаков, терминов, понятий.

Несколько облегчает задачу составления и оценивания ТЗ данного вида введение в перечень вариантов ответов дистрактора. Формулирование дистрактора проще, чем вариантов частично верных ответов, упрощается трактовка частично верных ответов на тестовое задание: важно, чтобы дистрактор в последовательности вариантов ответов находился на последней позиции, а верный ответ – на первой [6, 8]. В связи с этим ТЗ данного вида делят на ТЗ с градацией верных ответов, подчеркивая отсутствие среди вариантов ответа дистрактора, и ТЗ с градацией ответов, допуская или полагая, что среди вариантов ответа присутствует ложное, а не частично верное утверждение.

Однако испытуемые, не предупрежденные о наличии или отсутствии дистрактора в ТЗГО, допускают большее количество ошибок, чем уведомленные заранее, так как наиболее применимые по экспертным оценкам индивидуальные алгоритмы решения ТЗГО с дистрактором и без такового отличаются.

Пример 3.16. Корабли, снаряженные в первую экспедицию Колумба (1492–1493 гг.), назывались:

- 1) «Санта-Мария», «Пинта», «Нинья»;
- 2) «Санта-Анна», «Санта-Мария», «Нинья»;
- 3) «Санта-Мария», «Нинья».

Если испытуемый знает, что из трех ответов один неверный и один частично верный, то он будет выбирать неверный ответ из 1 и 2, т. е. угадывать с вероятностью 0,5 их позиции в последовательности, определяя за вариантом 3 вторую позицию в ответе на ТЗ.

Аналогичная ситуация имеет место и при ответе на ТЗМВ: если испытуемый заранее не предупрежден о том, что верный ответ может быть не единственным, вероятность его ошибки существенно возрастает [26, 75].

Добавление дистракторов в ТЗГО, в том числе за счет уменьшения количества частично верных ответов, обеспечивающее упрощение составления и оценивания задания, повышение его надежности, не упрощает решение проблемы оценивания частично верных ответов [6, 8, 38].

Тестовые задания с упорядочиванием дистракторов по структуре сходны с ТЗВО, а по формату решения близки к ТЗГО и ТЗУП. Последнее обстоятельство обеспечивает возможность размещения ТЗУД в компьютерных тестовых оболочках, допускающих формат ответа на ТЗ в виде последовательности цифр, при условии сопровождения во-

проса короткой инструкцией о формате ответа. В определенной степени ТЗУД можно рассматривать как разновидность ТЗГО с единственным верным ответом, а ТЗВО – как упрощенный вариант ТЗУД. Однако отбор содержания, методические подходы к формулированию ответов и заданий в целом, алгоритмы решения и критерии их оценивания для ТЗУД и ТЗГО принципиально отличаются. Несмотря на необходимость оценивания частично верных ответов в обоих случаях, интерпретация отличия ответов, даваемых испытуемыми, от эталонных последовательностей как характеристики полноты и глубины освоения учебного материала также различна для ТЗУД и ТЗГО. Наиболее убедительна такого рода интерпретация, основанная на непосредственном анализе решений заданий экспертами [42, 87]. Тем не менее уже сама корректность оценочных процедур и алгоритмов, обеспечивающая достоверность и информативность КТОЗУО, является залогом не только эффективности первичного анализа и сортировки результатов КУДО в системах мониторинга качества реализации образовательных программ, но и обратных связей в учебном процессе [28, 94].

Решение ТЗУД можно рассматривать как симметричное отображение решения разновидности ТЗГО с дистрактором: в первом случае один безусловно верный ответ и убывающая по степени близости от него последовательность дистракторов, во втором – дистрактор в конце последовательности, которая начинается абсолютно верным ответом и продолжается частично верными.

Тесты, составленные из ТЗУД и ТЗГО, могут использоваться и как вспомогательный оценочный инструмент. Так, один из авторов настоящей монографии использовал на устном экзамене тест из 10 ТЗУД. Если испытуемый менее, чем в половине заданий указывал верно первую позицию последовательности безотносительно прочих, то из его оценки ответов на вопросы билета вычитался один балл. Если экзаменуемый в более, чем 80% заданий указывал верно первую позицию последовательности и полностью верно выполнял не менее половины заданий, то к его оценке ответов на вопросы билета прибавлялся один балл.

Относительно редкое использование ТЗУП разработчиками тестов объясняется не столько трудоемкостью их разработки, сколько неопределенностью и технологической сложностью учета частично верных ответов, неприспособленностью интерфейсов компьютерных программных оболочек для создания тестов к размещению ТЗ усложненной структуры, ограниченностью ДЕСО, для проверки которых можно формулировать ТЗУП.

Практика применения ТЗУП позволила нам предложить их следующую систематизацию.

1. Тестовые задания на установление последовательности на основе сравнения известных количественных (измеренных или установленных) характеристик.

Пример 3.17. Расположите по мере удаленности от Москвы города: Санкт-Петербург; Казань; Нижний Новгород; Вологда; Воронеж.

Пример 3.17а. Расположите по мере удаленности от Москвы города: Томск; Омск; Екатеринбург; Казань; Тюмень.

Пример 3.18. Расположите города по убыванию численности населения: Казань; Санкт-Петербург; Архангельск; Владимир; Владивосток; Оренбург.

Пример 3.18а. Расположите города по убыванию численности населения: Казань; Ростов-на-Дону; Самара; Красноярск; Саратов; Омск.

Пример 3.18б. Расположите города по убыванию численности населения: Новосибирск; Архангельск; Тольятти; Сочи; Владимир; Балашиха.

Пример 3.19. Расположите по мере возрастания частоты: рентгеновское излучение; ультрафиолетовое излучение; инфракрасное излучение; гамма-излучение; оптическое излучение.

Пример 3.19а. Расположите по мере возрастания длины волны радиодиапазоны: КВЧ; УВЧ; СВЧ; ОВЧ; ГВЧ.

Трудность ТЗ может варьироваться, например, однонаправленностью и разнонаправленностью шкал измерения количественных параметров объектов, что приводит к сопоставлению таких параметров и формированию последовательностей на основе одномерных или квазидвумерных шкал (примеры 3.17 и 3.17а), величиной различий сравниваемых параметров (примеры 3.18 и 3.18б), сопоставлением объектов из различных категорий или включением в перечень сопоставления относительно малоизвестных объектов (примеры 3.18 и 3.18в), использованием аббревиатур (пример 3.19), сопоставлением производных или реже употребляемых для сравнения параметров (пример 3.19). Решение ТЗ данной разновидности требует конкретных знаний, запоминания численных величин или их соотношений.

2. Тестовые задания на установление последовательности на основе примерного (качественного) сравнения.

Пример 3.20. Расположите с запада на восток реки: Енисей; Обь; Печора; Лена; Северная Двина.

Пример 3.20а. Расположите с востока на запад устья рек: Дон; Иртыш; Кама; Волга; Обь.

Данная разновидность ТЗУП является упрощенной версией первой разновидности, сходной с примером 3.17а. Однако, в отличие от примера 3.17а, в примерах 3.20 и 3.20а сравниваются не количественные характеристики объектов, а их относительное (взаимное) расположение в четко указанном направлении (в примере 3.17а такое направление, облегчающее решение, еще следует определить). Трудность ТЗ в примере 3.20а повышена за счет его квазидвумерной формулировки, предполагающей вначале примерное определение по карте объектов сравнения.

Другими примерами, показывающими различия между первой и второй разновидностями, могут служить следующие ТЗУП.

Пример 3.21. Расположите события в хронологической последовательности: Куликовская битва; Грюнвальдская битва; битва при Молодях; «стояние на Угре»; основание Нижнего Новгорода.

Пример 3.21а. Укажите порядок правления российских императоров: Александр I, Павел I, Николай, Петр III, Анна I.

Повышение уровня трудности ТЗ может достигаться за счет расширения географии событий, т. е. обращения к истории различных государств, что требует от испытуемых умения сопрягать несколько хронологических шкал.

3. Тестовые задания на составление определений или корректных формулировок из отдельных слов.

Пример 3.22: my sister; the, of, aunt, father, is.

Пример 3.23: катет; на; прилежащего угла; есть; гипотенуза; косинус; произведение.

Пример 3.24: семеро; один; с сошкой; с ложкой.

Данную разновидность ТЗУП наиболее просто применять при контроле знаний по иностранным языкам.

Усложнение решения ТЗ может обеспечиваться вариативностью склонений существительных, прилагательных, числительных и местоимений и спряжений глаголов.

Пример 3.23а: катет (-а); на; прилежащий (-его); угол (угла); есть; гипотенуза (-ы); косинус (-а); произведение (-я).

Пример 3.25: один (одни); семеро (-ым); руками; пашет (-ут); машет (-ут).

Пример 3.26. Составьте формулировку закона Бойля-Мариотта: при постоянном (-ой); произведение (-и); давление (-и); температуры (-е); объема (-е); и; газа; неизменно (-ы).

4. Тестовые задания на запись формул из отдельных символов.

Пример 3.27. Составить формулу из символов: I ; R ; U ; $=$; $\times$.

При оценивании примеров 3.26 и 3.27, содержащих произведения величин, следует учитывать два возможных верных ответа, обусловленных изменением мест сомножителей. В то же время применение знака деления в записи формул или термина «отношение» в словесных формулировках также делает возможными два верных ответа.

Пример 3.28. Составьте формулу из символов: I ; R ; U ; $=$; $/$.

Верные ответы: $I = U / R$; $R = U / I$.

К единственному верному ответу приводит конкретизация задания: в первом случае – составить выражение для закона Ома для участка цепи; во втором – составить выражение для сопротивления участка электрической цепи. Уровень трудности ТЗ повышается путем увеличения количества переменных и особенно знаков арифметических действий или функций (тригонометрических, логарифмических, экспоненциальной и др.).

5. Тестовые задания на сопоставление результатов качественных задач.

Пример 3.29. Расположите фигуры по мере убывания отношения их периметра к площади: квадрат; прямоугольник; равносторонний треугольник; круг; равнобедренный (неравносторонний) треугольник; правильный шестиугольник.

Пример 3.30. Камень бросают вертикально вверх, и он падает на землю через время T . Расположите в порядке возрастания значения кинетической энергии камня в моменты времени: T ; $T/2$; $T/3$; $T/4$; $5T/6$.

6. Тестовые задания на сопоставление количественных результатов однотипных задач или серийных вычислений.

Пример 3.31. Ядро массой 5 кг вылетает из ствола стоящей на земле пушки под углом 30° к горизонту со скоростью 50 м/с. Расположите в порядке возрастания потенциальную энергию ядра в моменты времени $t = 1$ с; 1,5 с; 2,5 с; 3 с; 5 с. Высотой среза ствола относительно земли и сопротивлением воздуха пренебречь.

Важным свойством ТЗУП является их хорошая сочетаемость с ТЗ других видов – ТЗВО, ТЗМВ, ТЗУС [1, 6, 8, 22]. Использование парных ТЗ различных видов, удаленных друг от друга по порядку следования, позволяет проверять знание ДЕСО более глубоко и всесторонне, одновременно обеспечивая высокий уровень надежности теста. В качестве такой пары ТЗ можно рассмотреть примеры 3.18 и 3.32.

Пример 3.32. Рентгеновское излучение по-другому называется: альфа-излучением; бета-излучением; икс-лучами; реликтовым излучением; космическими лучами.

Ввиду сравнительно сложной структуры ТЗУП желательно, чтобы ТЗ другого вида имело более простую конструкцию. В то же время общая трудность решения пары ТЗ может быть выше трудности каждого задания в отдельности для испытуемого, слабо владеющего контролируемым материалом, и напротив, для испытуемого с хорошими знаниями решение одного из ТЗ может упростить решение другого ТЗ (примеры 3.33 и 3.34).

Пример 3.33. Укажите последовательность смены английских королевских династий: Йорки; Стюарты; Тюдоры; Ланкастеры; Вельфы.

Пример 3.34. Установите соответствие между правителями государств Европы (Леопольд I, Яков I, Георг I, Вильгельм II, Генрих VIII) и династиями (Гогенцоллерны, Габсбурги, Тюдоры, Стюарты, Вельфы (Ганноверская династия)).

Тестовые задания на установление последовательности при реализации их дидактического потенциала представляют собой эффективный инструмент оценивания знаний и умений в различных предметных областях. Применение ТЗУП в естественных науках и математике наделяет задания характерными особенностями, которые позволяют выделить их в отдельную разновидность данного вида ТЗ. Подобные различия, обуславливающие дополнительную классификацию ТЗУП, свойственны и заданиям, оперирующим количественными характеристиками объектов.

Разработка ТЗУП относительно трудоемка и требует от конструкторов тестов творческого подхода, обширных предметных знаний и скрупулезности, благодаря чему создатель ТЗ данного вида в процессе работы в определенной мере повышает свою квалификацию в предметной области, а продуманный и целенаправленный подбор элементов последовательностей обеспечивает реализацию не только контролирующей, но и обучающей функции ТЗУП. Не уступая ТЗВО в простоте редактирования, ТЗУП обладает значительно более высокой потенциальной устойчивостью к угадыванию верного ответа, что позволяет повысить надежность КТОЗУО за счет уменьшения вероятности ошибок второго рода, при этом вероятность ошибок первого рода при оценивании знаний и умений ТЗУП находится на одном из самых низких уровней среди всех видов ТЗ [26, 87].

3.3. Совершенствование оценочных шкал и алгоритмов начисления баллов за решение тестовых заданий

При конструировании КТОЗУО обучающихся ключевой задачей является согласование соответствующего положительной оценке порогового значения набранных баллов или диапазонов баллов оценочной шкалы с трудностью и надежностью теста. Для определения порогов и диапазонов баллов разработчик теста (при его безусловной валидности) должен сопоставить между собой распределение ТЗ по уровню трудности в их выборке (предварительно определив уровни трудности отдельных ТЗ с учетом вероятности угадывания правильных ответов на задания отдельных видов и уровней трудности), продолжительность тестирования, время, прошедшее от момента завершения формирования оцениваемых знаний и умений с возможным количеством набранных баллов. Важную роль при таком сопоставлении играет корректность целеполагания, выражающегося в направленности контроля (текущего, итогового, промежуточного), его дидактическом (степени обучающего воздействия) и воспитательном потенциале, применении полученных результатов (для выставления оценок по изученной дисциплине или части ее содержания, анализа оценок в системе внутреннего или внешнего мониторинга качества образования, целевого контроля в ходе проверочных или аттестационных мероприятий), а также прогноз ожидаемых результатов.

Для достижения максимальной достоверности оценок учебных достижений обучающихся, полученных путем компьютерного тестирования, в процессе разработки теста необходимо оперировать несколькими количественными оценками сопоставляемых или оптимизируемых параметров, которые могут быть получены на основании различных подходов:

а) эвристического, основанного на субъективных представлениях одного или нескольких специалистов, корректность которых определяется их опытом и достижениями в разработке и применении КТОЗУО, интерпретации и анализе результатов тестирования (наиболее корректной реализацией эвристического подхода является метод экспертного моделирования, применяемый для оценивания потенциальных возможностей обучающихся решать ТЗ различных уровней трудности и видов [40, 87], по результатам которого определяется структура выборки заданий теста, время тестирования и диапазоны баллов для пересчета в оценки);

б) статистического, позволяющего на основе статистической взаимосвязи уровня трудности ТЗ уровня подготовки тестируемых и вероятности правильного решения ТЗ, вычислять один из трех параметров по заранее известным двум другим (наиболее распространенной статистической моделью, используемой для этих целей, является модель Г. Раша [8, 40, 59, 74, 88]; как правило, на основе данной модели в зависимости от задачи исследования подтверждается или оценивается достоверность результатов тестирования, определяется трудность отдельных ТЗ или теста в целом, а также индивидуальный или коллективный уровень подготовки испытуемых);

в) формализованного, реализующего математические модели, в аналитическом виде связывающие между собой сопоставляемые параметры, что позволяет по набору известных величин, определяемых с высокой степенью точности или достоверности или задаваемых разработчиком теста, вычислять значения искомых величин, в том числе диапазонов баллов оценочных шкал [1, 59];

г) эмпирического, оперирующего данными проведенных ранее «пристрелочных» тестов, с дальнейшей пошаговой корректировкой диапазонов баллов оценочной шкалы, или уровнем трудности заданий, или относительными объемами тестируемых ДЕСО [8, 30, 59].

Каждый из этих подходов обладает очевидными достоинствами и недостатками. Основным преимуществом эвристического подхода является оперативность разработки КТОЗУО, однако достоверность результатов тестирования сильно зависит как от квалификации разработчика, в основном определяемой его практическим опытом преподавания и тестирования учебных достижений обучающихся, так и его интуиции. Получение экспертных оценок, способных исключить или минимизировать субъективные ошибки разработчика, зачастую затруднено невозможностью оперативно организовать экспертизу с достаточным количеством квалифицированных экспертов в предметной области, способных давать согласованные оценки.

Статистический подход относительно прост в реализации, а применение модели Раша является своего рода сертификацией КТОЗУО и результатов соответствующих контрольных мероприятий для значительного количества разработчиков и пользователей тестов. Однако практическая реализация данного подхода предполагает определение заранее (по результатам предыдущих контрольных мероприятий, в том числе по другим предметным областям) уровня подготовки испытуемых или уровня трудности ТЗ (например, по результатам тестирования другого контингента или на основе собственного

мнения разработчика теста либо экспертной оценки). Как правило, допущения в определении исходных данных модели, способные привести к неопределенности (расходимости) конечных результатов, на практике не учитываются. Применение модели Раша наиболее корректно для ТЗ, решение которых предполагает развернутый ответ, дополнение, численное значение, символьное преобразование, т. е. исключающих угадывание верного ответа из предлагаемых вариантов. Трудности ее применения возникают и в случае оценивания частично верных решений, в том числе в ТЗМВ, ТЗГО, ТЗУД, ТЗУС [1, 8, 26, 73].

В математических моделях, применяемых в формализованном подходе, используются параметры, определяемые эмпирическим или экспертным методом, что зачастую не позволяет использовать модели, разработанные для определенных дидактических условий (контингент, вид контроля, содержание контроля), в других условиях без дополнительной «настройки» КТОЗУО. Применяемый математический аппарат, несмотря на его относительную простоту, вызывает затруднения в освоении разработчиками тестов в гуманитарных предметных областях. В равной степени эта проблема относится и к статистическому подходу. Основными достоинствами формализованного подхода являются однозначность и наглядность зависимостей различных параметров моделей, относительная простота интерпретации и анализа результатов тестирования, корректировки тестов при изменении условий обучения или характеристик контингента испытуемых, встраиваемость моделей в алгоритмы мониторинга внутренней системы качества образования [1, 8, 30, 59].

Эмпирический подход может опираться на статистические методы обработки данных, а также математические модели, например, оценки надежности ТЗ и тестов в целом, однако чаще всего реализуется в эмпирико-эвристическом методе, когда в основе корректировки тестов или оценочных шкал лежит субъективное мнение разработчика. В достижении приемлемой достоверности оценок квалификация разработчика сказывается на количестве итераций (процедур «пристрелочного» тестирования), умении корректно применять статистические методики обработки небольших массивов данных. Как правило, разработка КТОЗУО на основе эмпирического подхода требует большего времени относительно других подходов. Еще одним его недостатком является чувствительность получаемых оценочных шкал к изменению условий обучения, требующих переработки тестов или корректировки значений диапазонов баллов, соответствующих принятой оценочной шкале [1, 30, 59, 74].

Недостатки, присущие каждому подходу в отдельности, приводят к тому, что на практике разработчики КТОЗУО используют комбинированные подходы, компенсируя недостатки какого-либо взятого ими за основу подхода или метода на его основе элементами, заимствованными из другого подхода. Так, экспертные оценки или экспертное моделирование применяются в реализации статистического и формализованного подходов; статистический подход может использоваться для оценки достоверности результатов применения тестов независимо от подходов, положенных в основу их разработки; формализованный подход позволяет по результатам КУДО выявлять и учитывать ошибки в составлении тестов [89].

Оценочная шкала в процентах от максимального количества набранных баллов может применяться в ходе любых видов аттестационных или контрольных мероприятий.

Наиболее просто определить нижний порог отличной оценки, хотя, как показывает анализ тестов, их разработчики далеко не единодушны во мнениях относительно данного показателя, указывая его в пределах от 80% (реже 75%) до 90% (реже 95%). Помимо субъективизма разработчиков, на данные оценки влияет уровень трудности ТЗ и вид контроля: текущий, промежуточный, остаточных знаний.

Исходя из того, что «настоящий отличник» должен уметь правильно выполнить 100% ТЗ, учтем в конечном значении нижнего порога отличной оценки возможные нюансы как самого теста, так и оценочной процедуры.

Будем полагать, что в тестовой выборке из N ТЗ представлены задания трех уровней трудности:

$$N = N_{\text{пр}} + N_{\text{ср}} + N_{\text{тр}} \equiv N \cdot (\alpha_{\text{пр}} + \alpha_{\text{ср}} + \alpha_{\text{тр}}), \quad (3.1)$$

где $N_{\text{пр}}, N_{\text{ср}}, N_{\text{тр}}$ – соответственно количество простых, средней трудности и трудных ТЗ в тестовой выборке, $\alpha_{\text{пр}}, \alpha_{\text{ср}}, \alpha_{\text{тр}}$ – доли в выборке соответственно простых, средней трудности и трудных ТЗ:

$$\alpha_{\text{пр}} + \alpha_{\text{ср}} + \alpha_{\text{тр}} = 1. \quad (3.2)$$

По результатам экспертного моделирования [87], отличник решает правильно все ТЗ, за исключением некоторого количества $N_{\text{ош тр}}$, которое определяется количеством ТЗ повышенной трудности, в том числе необходимых для контроля деградации КТОЗУО [29].

Количество трудных ТЗ, решенных «настоящим отличником»:

$$N_{\text{р тр}} = N_{\text{тр}} - N_{\text{ош тр}}, \quad (3.3)$$

а экспертная оценка их доли в общем количестве:

$$\alpha_{\text{тр}}^* = \frac{N_{\text{р тр}}}{N_{\text{тр}}} \cdot \alpha_{\text{тр}} \equiv \beta_{\text{тр}} \cdot \alpha_{\text{тр}}. \quad (3.4)$$

Тогда для «настоящего отличника»:

$$N_{\text{р}} = N_{\text{р пр}} + N_{\text{р ср}} + N_{\text{р тр}} = N_{\text{пр}} + N_{\text{ср}} + N_{\text{тр}} = N \cdot (\alpha_{\text{пр}} + \alpha_{\text{ср}} + \alpha_{\text{тр}}^*); \quad (3.5)$$

$$\alpha_{\text{пр}}^* = \alpha_{\text{пр}}; \alpha_{\text{ср}}^* = \alpha_{\text{ср}}.$$

Учтем вероятность ошибок первого рода, когда испытуемый, обладая достаточными знаниями и умениями, неверно выполняет соответствующее ТЗ. Тогда для «настоящего отличника» количество решенных ТЗ:

$$N_{\text{р}} = N_{\text{р пр}} - N_{\text{пр I}} + N_{\text{р ср}} - N_{\text{ср I}} + N_{\text{р тр}} - N_{\text{тр I}}, \quad (3.6a)$$

или

$$N_{\text{р}} = N_{\text{пр}} - N_{\text{пр I}} + N_{\text{ср}} - N_{\text{ср I}} + N_{\text{тр}} - N_{\text{тр I}}, \quad (3.6b)$$

где $N_{\text{пр I}}$, $N_{\text{ср I}}$ и $N_{\text{тр I}}$ – количество ошибок первого рода, допускаемых «настоящим отличником» при решении простых, средней трудности и трудных заданий соответственно.

Снижение порога отличной оценки с учетом ошибок первого рода, обусловленных волнением, краткосрочной потерей внимания или концентрации, сомнением в правильности формулировок ТЗ при невозможности их уточнить, по нашему мнению, необходимо, так как целью КТОЗУО является проверка знаний и умений испытуемых, а не их внимания или стрессоустойчивости. Тем самым в некоторой степени нивелируется различие компьютерного тестирования на основе неадаптивных тестов и собеседования с преподавателем в части невозможности окончательно выявить освоение или неосвоение конкретной ДЕСО с помощью уточняющих вопросов и т. п.

Доли решенных ТЗ «настоящим отличником» с учетом вероятностей $\rho_{\text{пр I}}$, $\rho_{\text{ср I}}$ и $\rho_{\text{тр I}}$ ошибок первого рода:

$$\alpha_{\text{пр}}^* = \alpha_{\text{пр}} \cdot (1 - \rho_{\text{пр I}}); \quad (3.7a)$$

$$\alpha_{\text{ср}}^* = \alpha_{\text{ср}} \cdot (1 - \rho_{\text{ср I}}); \quad (3.7b)$$

$$\alpha_{\text{тр}}^* = \alpha_{\text{тр}} \cdot (1 - \rho_{\text{тр I}}). \quad (3.7b)$$

Порог отличной оценки, равный проценту решенных «настоящим отличником» ТЗ с учетом ошибок первого рода:

$$P_{\text{отл}} = (\alpha_{\text{пр}}^* + \alpha_{\text{ср}}^* + \alpha_{\text{тр}}^*) \cdot 100\% =$$

$$= [\alpha_{\text{пр}} \cdot (1 - \rho_{\text{пр I}}) + \alpha_{\text{ср}} \cdot (1 - \rho_{\text{ср I}}) + \beta_{\text{тр}} \cdot \alpha_{\text{тр}} \cdot (1 - \rho_{\text{тр I}})] \cdot 100\%. \quad (3.8)$$

Применительно к упрощенной модели «хорошиста» следует полагать, что он решает все простые ТЗ без учета ошибок первого рода и все ТЗ средней трудности. На практике решение реальными «хорошистами» отдельных трудных заданий компенсируется нерешением

отдельных ТЗ средней трудности, относящихся к сравнительно слабо освоенным ДЕСО.

Поскольку количество нерешенных ТЗ в упрощенной модели «хорошиста» $N_{\text{ош}} = N_{\text{тр}}$ может заметно отличаться от количества ТЗ, нерешенных «настоящим отличником» $N_{\text{ош тр}}$, при определении порогового значения оценки «хорошо» следует учесть угадывание испытуемым верных ответов на трудные ТЗ:

$$N_{\text{уг тр}} = \rho_{\text{тр II}} \cdot (N_{\text{тр}} - N_{\text{р тр}}),$$

где $N_{\text{уг тр}}$ – среднее количество угаданных ответов на трудные ТЗ, $\rho_{\text{тр II}}$ – средняя вероятность угадывания ответов (средняя вероятность ошибки второго рода) на трудные ТЗ с добавлением полученного значения в выражения для расчета решенных ТЗ, аналогичные (3.6а) и (3.6б). Однако, как показывает практика разработки и применения КТОЗУО, вклад $N_{\text{уг тр}}$ в общее количество решенных ТЗ настолько мал, что им можно пренебречь. Тогда нижний порог оценки «хорошо» с учетом ошибок первого рода:

$$P_{\text{хор}} = (\alpha_{\text{пр}}^* + \alpha_{\text{ср}}^*) \cdot 100\% = [\alpha_{\text{пр}} \cdot (1 - \rho_{\text{пр I}}) + \alpha_{\text{ср}} \cdot (1 - \rho_{\text{ср I}})] \cdot 100\%. \quad (3.9)$$

Установление порога оценки «удовлетворительно» характеризуется наибольшей степенью субъективизма, так как частные критерии такой оценки, которыми руководствуются в различных образовательных организациях, допускают неосвоение отдельных ДЕСО, что находит отражение и в КТОЗУО.

Как правило, для получения оценки «удовлетворительно» недостаточно решить все простые ТЗ или их существенно большую часть:

$$N_{\text{р}} = N_{\text{р пр}} + N_{\text{р ср}} = N_{\text{пр}} - N_{\text{ош пр}} + N_{\text{ср}} - N_{\text{ош ср}}, \quad (3.10)$$

где $N_{\text{ош пр}}$ и $N_{\text{ош ср}}$ – количество нерешенных ТЗ, простых и средней трудности соответственно.

Доля решенных ТЗ в этом случае:

$$\frac{N_{\text{р}}}{N} = \alpha_{\text{пр}}^* + \alpha_{\text{ср}}^* = \frac{N_{\text{р пр}}}{N_{\text{пр}}} \cdot \alpha_{\text{пр}} + \frac{N_{\text{р ср}}}{N_{\text{ср}}} \cdot \alpha_{\text{ср}} \equiv \beta_{\text{пр}} \cdot \alpha_{\text{пр}} + \beta_{\text{ср}} \cdot \alpha_{\text{ср}}. \quad (3.11)$$

С учетом ошибок первого рода и угадывания верных ответов:

$$N_{\text{р}} = N_{\text{р пр}} - N_{\text{пр I}} + N_{\text{уг пр}} + N_{\text{р ср}} - N_{\text{ср I}} + N_{\text{уг ср}}, \quad (3.12)$$

где $N_{\text{уг пр}} = \rho_{\text{пр II}} \cdot (N_{\text{пр}} - N_{\text{р пр}})$ и $N_{\text{уг ср}} = \rho_{\text{ср II}} \cdot (N_{\text{ср}} - N_{\text{р ср}})$ – количество угаданных правильных ответов соответственно на простые и средней трудности ТЗ, $\rho_{\text{пр II}}$ и $\rho_{\text{ср II}}$ – соответственно средние вероятности угадывания решения (ошибок второго рода) простых и средней трудности ТЗ.

Выражения для расчета порога оценки «удовлетворительно» будет отличаться от выражения (3.9) учетом доли самостоятельно решенных заданий и вероятности угадывания верных ответов:

$$P_{уд} = [\alpha_{пр} \cdot \beta_{пр} \cdot (1 - \rho_{пр I} + \rho_{пр II}) + \alpha_{ср} \cdot \beta_{ср} \cdot (1 - \rho_{ср I} + \rho_{ср II})] \cdot 100\%. \quad (3.13)$$

Проблемой применения выражения (3.13) является то, что вероятность угадывания верных ответов ТЗ испытуемыми с различными уровнями подготовленности различна. Например, при решении ТЗВО тестируемые с более высоким уровнем подготовки могут исключить из количества угадываемых вариантов ответа большее количество дистракторов, сведя это количество до двух-трех и повысив тем самым среднюю вероятность угадывания примерно до 0,4 [26, 29]. В то же время обучающиеся, по уровню подготовки отнесенные к «троечникам», характеризуются наиболее сильной дифференциацией знаний по сравнению с категориями более успешных обучающихся: от постоянно неуспевающих до «почти «хорошистов». Это же обстоятельство приводит к тому, что для слабо подготовленных испытуемых возрастает вероятность ошибки первого рода. Например, испытуемый может не воспринять незнакомый ему термин, используемый при формулировании ТЗ и являющийся эквивалентом известного ему термина. В связи с этим рекомендуется трактовать возможные дополнительные ошибки первого и второго рода в пользу испытуемых, т. е. определять в выражениях (3.8), (3.9) и (3.13) значения вероятностей ошибок первого рода по верхней границе их оцениваемого диапазона, а значения вероятностей ошибок второго рода в выражении (3.13) – по нижней границе.

При использовании КТОЗУО важно накапливать, обобщать и анализировать эмпирический материал, сопоставлять вносимые в тесты коррективы с изменением достоверности получаемых оценок на возможно большем объеме данных с учетом задач тестирования, контингента, предметных областей и иных дидактических условий. Это не только обеспечивает корректность применения статистических методов обработки полученных данных, но и приводит к постепенному вытеснению как из уже используемых КТОЗУО, так и из процедур разработки новых тестов эвристической компоненты и замещению ее статистическими или явными математическими моделями.

Оценивание частично верных ответов представляет собой задачу с множеством решений, выбор которых осуществляется разработчиками КТОЗУО исходя из трудности ТЗ, целей тестирования, значения порога положительной оценки или диапазонов оценочной шкалы и собственных трактовок неполных или частично правильных решений.

Начисление баллов за частично верные ответы основывается на одном из следующих подходов (принципов).

I. Частично верный или неполный ответ приравнивается к неверному ответу и так же оценивается. Таким подходом наиболее часто пользуются при оценивании ТЗМВ, решение которых представляет собой полную совокупность верных ответов из предлагаемого испытуемому набора верных ответов и дистракторов.

II. За верную часть решения начисляются баллы пропорционально отношению объема этой части к общему объему задания. Такой подход зачастую применяется при оценивании ТЗУС и ТЗМВ [26].

III. За верную часть решения начисляются баллы в соответствии с ее относительным объемом и статусом. Так, в алгоритмах начисления баллов за решение ТЗМВ и ТЗУС целесообразно учитывать степень важности верных вариантов ответов, если таковые возможно разделить по данному признаку (например, на основные и дополнительные). В ТЗГО такое разделение определяется самим характером задания, в ТЗУД подобным образом (по степени близости к верному ответу) разделяются дистракторы, а наивысшим статусом обладает верный ответ [30].

IV. За неверно решенную часть ТЗ начисляются штрафы в соответствии с ее относительным объемом. Данный подход является антиподом подхода, изложенного в пункте II, и реализуется в тех случаях, когда он обеспечивает упрощение вычислительных процедур или наглядность ошибок.

V. За неверно решенную часть ТЗ начисляются штрафы, учитывающие как ее относительный объем, так и статус.

VI. За верно решенную часть задания начисляются баллы, за неверно решенную часть – штрафы в равных пропорциях относительно общего объема задания. Такой подход наиболее оправдан при оценивании ТЗУС и ТЗМВ для того, чтобы разделить случаи выбора неверного варианта ответа (неверного решения) и отказа от выбора (отсутствия решения) [30, 81].

VII. Модификация подхода, изложенного в пункте VI, отличающаяся тем, что штрафы начисляются как за неверно решенную, так и за нерешенную часть ТЗ.

VIII. Модификация подхода, изложенного в пунктах IV, VI и VII, отличающаяся тем, что коэффициенты начисляемых баллов или штрафов различаются в зависимости от того, насколько разработчик КТОЗУО стремится подчеркнуть важность верных решений (частичного знания) или ошибок (частичного незнания). В частности, величи-

на штрафа может зависеть от степени отклонения неверно указанной позиции последовательности в ответе испытуемого от позиции в эталонной последовательности (ТЗУП, ТЗУД, ТЗГО).

IX. Баллы за верно решенную часть ТЗ и штрафы за его неверно решенную (и нерешенную) часть начисляются с учетом объемов, трудности и статусов соответствующих частей ТЗ, а также трактовок разработчиком КТОЗУО понятия «частично верное решение». Данный подход является наиболее сложно реализуемым, однако обеспечивает наиболее детализированную оценку ТЗ, так как позволяет осуществлять учет многих особенностей задания, характеризуется наличием настроек в вычислительных алгоритмах. Его реализация наиболее востребована в оценках ТЗГО, ТЗУД, а также ТЗМВ, в которых верные варианты ответа характеризуются различной значимостью [8, 30, 74, 90].

Как и для других ТЗ с усложненной структурой, для большинства разновидностей ТЗУП характерна возможность получения частично верных ответов, однако, в отличие от других видов ТЗ с усложненной структурой, оценивание частично верных решений ТЗУП осуществляется относительно простыми алгоритмами и расчетными формулами с минимальным количеством настраиваемых параметров [1, 30, 81].

Выбор подхода к оцениванию ТЗУП должен учитывать признаки представленных выше шести разновидностей ТЗ данного вида. Так, при установлении последовательностей количественных характеристик или качественного сравнения (см. приведенные в разделе 3.2 разновидности ТЗУП 1, 2) могут использоваться оценочные алгоритмы на основе подходов II, IV, VI, VIII, т. е. начисление балла за верное указание позиции в последовательности и (или) начисление штрафа за неверное указание позиции.

Значение начисленного балла определяется:

$$B_B = \frac{B_{max} \cdot N_B}{N}, \quad (3.14)$$

где B_{max} – максимальный начисляемый балл за полностью верное решение ТЗУП; N – количество элементов последовательности; N_B – количество верно указанных позиций элементов последовательности.

Значение штрафа за неверное указание позиций элементов последовательности:

$$Ш_H = \frac{Ш_{max} \cdot N_H}{N}, \quad (3.15)$$

где $\text{Ш}_{\max}$ – значение максимального штрафа; $N_{\text{н}}$ – количество неверно указанных позиций элементов последовательности.

Итоговый балл за выполнение ТЗ:

$$B_{\text{и}} = K_{\text{т}} \cdot (B_{\text{в}} - \text{Ш}_{\text{н}}), \quad (3.16)$$

при условии, если $B_{\text{в}} < \text{Ш}_{\text{н}}$. то $B_{\text{и}} = 0$.

$K_{\text{т}}$ в формуле (3.16) – коэффициент трудности или значимости задания.

Формулы (3.14) – (3.16) позволяют разработчику ТЗ проводить определенную оценочную политику, задавая соотношения между значениями максимально начисляемых баллов и максимальных штрафов, тем самым поощряя испытуемых за частично верные ответы и фрагментарные знания или поднимая планку требований.

Несмотря на простоту указанных подходов к оцениванию и их реализации, итоговый балл за выполнение ТЗ может зависеть от трактовок разработчиком или экспертами частично верного решения [6, 87], их интерпретации части верно указанных позиций последовательности.

Рассмотрим приведенный в разделе 3.2 пример 3.17а. Пусть верному решению (Казань, Екатеринбург, Тюмень, Омск, Томск) соответствует начисляемый балл, равный единице, т. е. за каждое верное указание позиции города в их последовательности испытуемый получает 0,2 балла.

Допустим, испытуемый указал в ответе следующую последовательность: Казань, Екатеринбург, Омск, Томск, Тюмень.

Формально в данной последовательности верно указаны две позиции, неверно – три. На основе подхода II ТЗ будет оценено в 0,4 балла.

Если полагать величины штрафов и начислений за одну соответственно неверно или верно указанную позицию последовательности равными (в данном примере 0,2 балла), то оценка ТЗ на основе подхода VI в соответствии с выражениями (3.14) – (3.16) составит 0 баллов.

При использовании подхода IV к оцениванию ТЗУП, по нашему мнению, следует руководствоваться принципом: ответ в виде последовательности, большая часть которой не совпадает с эталонной, не должен оцениваться положительно. Следовательно, для последовательности из пяти элементов

$$\text{Ш}_{\max} \geq B_{\max} \cdot \frac{5}{3},$$

а в общем виде для нечетного N

$$\mathcal{Ш}_{max} \geq \mathcal{Б}_{max} \cdot \frac{2N}{N+1},$$

для нечетного N

$$\mathcal{Ш}_{max} \geq 2\mathcal{Б}_{max}.$$

Однако рассматриваемый частично верный ответ может трактоваться как полученный вследствие единственной ошибки – размещения «Тюмень» не на третьей, а на пятой позиции, что обусловило сдвиг «Омск» и «Томск» на одну позицию влево. Для сравнения рассмотрим последовательность: Казань, Екатеринбург, Томск, Омск, Тюмень. Формально с эталонной последовательностью не совпадают две позиции, по подходу II решение ТЗ будет оценено в 0,6 балла, по подходу IV с выбором минимального штрафа – в 0,33 балла, по подходу VI – в 0,2 балла, т. е. все рассматриваемые подходы к оцениванию дают более высокие баллы. Вместе с тем в качестве причины неверного формирования последовательности может по-прежнему усматриваться единственная ошибка – перестановка мест «Томск» и «Тюмень», а расположение городов «Томск, Омск, Тюмень» выглядит даже в большей степени ошибочным, чем «Омск, Томск, Тюмень».

Данное кажущееся противоречие может быть устранено корректировкой формулировки задания, переводящей его из разновидности 1 ТЗУП в разновидность 2 (раздел 3.2).

Пример 176. Расположите с запада на восток города: Томск, Омск, Екатеринбург, Казань, Тюмень.

Выполнение данного задания предполагает попарное сравнение объектов по критериальному параметру «восточнее – западнее». Тогда последовательность «Казань, Екатеринбург, Омск, Томск, Тюмень» будет выглядеть более корректным ответом в сравнении с последовательностью «Казань, Екатеринбург, Томск, Омск, Тюмень» и, соответственно, оцениваться более высокими баллами на основании подходов II, IV, VI.

Применение подхода VIII в его наиболее простой реализации (начисление штрафа пропорционально величине отклонения указанной в ответе позиции от ее правильного местоположения) формализуется в виде:

$$\mathcal{Ш}_H = \sum_{i=1}^{N_H} N_i = \sum_{i=1}^{N_H} \Delta \mathcal{Ш} \cdot |n o_i - n \varepsilon_i| = \sum_{i=1}^N \Delta \mathcal{Ш} \cdot |O_i - \varepsilon_i|, \quad (3.17)$$

где $\Delta\Pi$ – минимальное значение штрафа, назначаемого за отклонение элемента последовательности в ответе на одну позицию от правильного местоположения; $no_i, nэ_i$ – соответственно номера позиций i -го элемента последовательности в ответе испытуемого и правильном решении ТЗ.

Тогда значение штрафа как за ответ «Казань, Екатеринбург, Омск, Томск, Тюмень», так и за ответ «Казань, Екатеринбург, Томск, Омск, Тюмень» в примере 1а при минимальном штрафе $\Delta\Pi = 1/N = 0,2$ составит: $\Pi_H = 0,8$.

Третья разновидность ТЗУП из их представленного в разделе 3.2 перечня, как правило, не предполагает частично верного ответа, и любое отклонение от эталонной последовательности приводит к нулевой оценке.

Для четвертой разновидности данного вида ТЗ в качестве частично верного ответа можно рассматривать, например, правильный выбор искомой величины и ее помещение в левую часть уравнения (пример 1.17 в формулировке: «Составить выражение для закона Ома для участка цепи»), правильный выбор функциональных зависимостей при ошибке с указанием аргументов функций или наоборот. Однако выявление такого рода частично правильных решений представляет интерес не с точки зрения оценивания знаний (так как ТЗ фактически не выполнено и не может быть оценено положительно), а в диагностическом аспекте – для определения степени усвоения или забывания материала, дополнительного дифференцирования слабо успевающих обучающихся, детализации содержания и методик индивидуального обучения и др.

Ввиду необходимости использования формул и выполнения символьных преобразований, в том числе в уме, при решении ТЗУП пятой разновидности сравнительно много ответов на такие задания содержат ошибки, т. е. оказываются частично верными. Наиболее универсальным основанием создания расчетных формул и алгоритмов для оценивания таких ответов является подход VIII, позволяющий учитывать степень несовпадения сгенерированного испытуемым ответа с эталонной последовательностью.

Подходы II, IV, VI применимы для оценивания решений ТЗУП пятой разновидности упрощенными алгоритмами начисления баллов и (или) штрафов.

При решении ТЗУП шестой разновидности частично верные решения могут быть обусловлены в том числе вычислительными ошибками системного и случайного характера. О случайных ошибках

свидетельствует «выпадение» какого-либо результата из в целом корректно сформированной последовательности, о системных – несовпадении всех позиций испытуемого в ответе и эталонной последовательности, численно характеризующееся аддитивной (отражающей сдвиг одной последовательности относительно другой) или мультипликативной (монотонно возрастающей или убывающей от элемента к элементу последовательности) величиной.

Для выявления и учета таких ошибок оценки за выполнение ТЗ, полученные на основе подхода VIII как основного, могут сравниваться с оценками, рассчитанными с применением формулы (3.15) при условии сопоставимости величин максимальных штрафов.

Подходы, обозначенные нами как III, V, IX, могут быть применены для оценивания ТЗУП в тех случаях, когда необходимы особые штрафные санкции за незнание основ, базовых понятий, отсутствие элементарной эрудиции. В соответствующих оценочных алгоритмах следует предусмотреть повышенный штраф за неправильное указание первой позиции последовательности, на которой могут, например, располагаться Джомолунгма – высочайшая вершина мира; Тихий океан – крупнейший из океанов; Байкал – глубочайшее озеро мира; водород – элемент с наименьшей атомной массой; Гагарин – первый человек, совершивший полет в космос; белый медведь – крупнейший сухопутный хищник; Юпитер – самая большая планета Солнечной системы и т. п.

Напротив, особый поощрительный статус, влекущий начисление повышенных баллов, может быть присвоен разработчиком ТЗ позиции последовательности, правильное указание которой испытуемым либо не может основываться на логически обоснованных закономерностях, либо свидетельствует о его широкой предметной эрудиции. К таковым, например, можно отнести: гелий – газ с самой низкой температурой замерзания; «Фредерик» – первый построенный в России военный корабль; Ляпидевский – первый Герой Советского Союза; Данилов – русский оружейник, артиллерист, изобретатель гаубицы «единорог» (1757 г.); «Ермак» – первый российский ледокол и др.

Тестовые задания с упорядочиванием дистракторов характеризуются наименьшей вероятностью ошибок второго рода, наивысшим дидактическим потенциалом, относительно просто редактируются, а изменение порядка дистракторов в ТЗ при каждом новом запуске КТОЗУО обеспечивает высокую надежность тестов [26, 30, 56]. Решение ТЗУД требует системного понимания сформулированной задачи, анализа вариантов ответа на предмет выявления их общего признака и оценки степени его наличия в каждом варианте, что существенно

затрудняет использование чатов ИИ и удаленных подсказок для получения верных ответов, выражаемых числовыми последовательностями. Несмотря на сложность составления ТЗУД, они представляются нам наиболее перспективной основой разработки «интеллектуальных» КТОЗУО.

Разработчики ТЗУД, как и многих ТЗ с усложненной структурой, сталкиваются с неоднозначностью начисления баллов за частично верное решение. В [56] приведен простой алгоритм начисления баллов за ТЗУД: размещение верного ответа на первой позиции – 1,0 балла, правильное размещение дистракторов на позиции со второй по пятую – 0,4, 0,3, 0,2 и 0,1 балла соответственно, за неправильное указание позиций баллы не начисляются. Там же рассмотрен вариант реализации штрафов за неверный выбор позиции – вычитание балла, равного начисляемому, из общей суммы (если сумма окажется отрицательной, ТЗ оценивается в 0 баллов). Однако реализация данных алгоритмов на практике, как полагают их пользователи, приводит к недостоверному оцениванию следующих казусов:

а) испытуемый ставит верный ответ (из пяти вариантов) на первую позицию, однако за счет неверного расположения оставшихся ответов получает 0 баллов за ТЗ; такой же результат получается и при любых вариантах отсутствия верного ответа на первой позиции вплоть до полного несовпадения позиций ответов с эталонной последовательностью;

б) 0,6 балла за ТЗ может быть получено при совпадении как лишь первой и третьей позиций (т. е. постановке на четвертую или даже вторую позиции дистрактора в заведомо абсурдной формулировке), так и лишь первой, четвертой и пятой позиций сформированной испытуемым и эталонной последовательностей;

в) 0,8 балла за ТЗ может быть получено при совпадении с эталонной последовательностью как лишь первой и второй позиций (т. е. постановке на третью или четвертую позиции заведомо абсурдного дистрактора), так и лишь первой, третьей и пятой позиций сформированной последовательности;

г) 1,0 балла за ТЗ может быть начислено при совпадении с эталонной последовательностью первой, второй и пятой позиций или первой, третьей, и четвертой позиций сформированной последовательности (с постановкой на четвертую или даже вторую позиции заведомо абсурдного варианта ответа и пяти предложенных).

Таким образом, ТЗУД, обеспечивая минимальную вероятность угадывания верного ответа и получения испытуемым незаслуженной

положительной оценки, нуждается в таких алгоритмах начисления баллов за решение ТЗ, которые бы исключали необоснованное преодоление порога положительной оценки КТОЗУО в «пограничных» ситуациях и способствовали детальному анализу итоговых результатов тестирования. Предпосылки создания таких алгоритмов обусловлены таким преимуществом ТЗУД по сравнению с другими видами ТЗ, как возможность дифференцировать оценки в зависимости от дидактических целей теста и глубины контроля учебных достижений [30, 56, 90].

Пример 3.35. Колумб вернулся из своей первой экспедиции (1492–1493 гг.) на корабле:

- 1) «Санта-Мария»;
- 2) «Пинта»;
- 3) «Нинья»;
- 4) «Тринидад»;
- 5) «Сокол тысячелетия».

Верный ответ: 3 2 1 4 5. Каравелла «Пинта» также достигла берегов Испании. «Санта-Мария» была в составе первой экспедиции Колумба, но затонула. «Тринидад» – название флагманского корабля первой кругосветной экспедиции Магеллана. «Сокол тысячелетия» – космический корабль Хана Соло, персонажа фантастического романа «Звездные войны».

Пример 3.36. Два положительных заряда:

- 1) притягиваются с силой, прямо пропорциональной расстоянию между ними;
- 2) притягиваются с силой, обратно пропорциональной квадрату расстояния между ними;
- 3) отталкиваются с силой, прямо пропорциональной расстоянию между ними;
- 4) отталкиваются с силой, обратно пропорциональной расстоянию между ними;
- 5) отталкиваются с силой, обратно пропорциональной квадрату расстояния между ними.

Верный ответ: 5 4 3 2 1. Критерий грубой ошибки: одноименные заряды не притягиваются, а отталкиваются. Второй по значимости критерий грубой ошибки: сила взаимодействия зарядов убывает по мере удаления зарядов друг от друга, а не наоборот. Таким образом, наименее грубой ошибкой является вариант ответа 4.

В отличие от ТЗМВ или ТЗУП, вопрос о необходимости ненулевого оценивания частично верных ответов ТЗУД не возникает.

Во-первых, подлежит хотя бы минимальной положительной оценке правильное размещение на первой позиции номера верного варианта ответа. Во-вторых, методически верно составленное ТЗУД позволяет выявить системность и комплексность знаний тестируемых, а обнуление оценки за ошибки, имеющие второстепенное значение, исключает возможность анализа усвоения обучающимися содержания обучения, подлежащего проверке. В-третьих, ТЗУД позволяют сделать максимально компактным КТОЗУО за счет высокой надежности заданий и возможности комплексной проверки знаний, но обнуление частично верных ответов при сравнительно малом количестве заданий приводит к снижению прецизионности оценочной шкалы и дифференцирующей способности теста [24, 26, 29, 43].

Оценивание частично верных решений ТЗ сводится к присвоению некоторых положительных числовых значений (в формате десятичной дроби) частично верным вариантам ответа и (или) начислению штрафов – отрицательных числовых значений при указании неверных вариантов ответов на позициях верных. Начисление или неначисление штрафов определяет две группы алгоритмов оценивания [6, 38].

Особенностью оценивания ТЗУД является учет грубых ошибок [26, 56]. Мнения специалистов о приоритетах – верный вариант ответа на первой позиции или нелепый вариант ответа на последней позиции – расходятся. С одной стороны, знание верного ответа важно, с другой – указание крайне неверного ответа на позиции, близкой к верному, свидетельствует о поверхностном знании тестируемым данного вопроса. Что является более приемлемым в решении примера 3.35: указание «Сокола тысячелетия» на второй позиции вслед за «Ниньей» или первая и вторая позиции «Пинта» и «Нинья» соответственно и последняя позиция «Сокола тысячелетия»? Различие мнений на этот счет отражается в разнообразии методик и алгоритмов начисления баллов за ТЗУД.

Важным свойством алгоритмов оценивания ТЗУД с начислением штрафов является возможность регулировки начисляемых баллов и штрафов. Пусть «штрафной» компонент имеет вид [90]:

$$P_i = \begin{cases} \alpha_i \cdot \beta_i^{(|i-n|-\delta_i) \cdot \gamma_i}; & i \neq n \\ 0; & i = n \end{cases}, \quad (3.18)$$

где $i = 2, 3, 4 \dots N$ – номер позиции дистрактора в правильном ответе на ТЗУД в целом в виде последовательности цифр – номеров позиций (эталонной последовательности); для верного варианта ответа на вопрос $i = 1$; n – номер позиции дистрактора или верного варианта ответа на вопрос в указанной тестируемым последовательности цифр;

N – количество вариантов ответа на вопрос тестового задания; $\alpha_i, \beta_i, \gamma_i, \delta_i$ – параметры, задаваемые при настройке программы тестирования.

Значения $\alpha_i, \beta_i, \gamma_i, \delta_i$ задаются конструктором теста исходя из его трактовки значимости характера тех или иных ошибок.

Очевидно, условие

$$\sum_{i=1}^N P_i \geq 1 \quad (3.19)$$

позволяет формировать оценку B_1 за ТЗУД в формате десятичной дроби только штрафными баллами, за счет их вычитания из 1:

$$B_1 = 1 - \sum_{i=1}^N P_i, \quad (3.20)$$

$$B_1 = 0, \text{ если } \sum_{i=1}^N P_i > 1. \quad (3.20a)$$

Тогда оценка за ТЗУД по Т-балльной шкале:

$$B_T = B_1 \cdot T. \quad (3.21)$$

Рассмотрим применение данной методики на примере 3.35.

Пусть для любого i : $\alpha_i = 0,1$, $\beta_i = 2$, $\gamma_i = 1$, $\delta_i = 1$. При вводе последовательности 1 3 4 2 5 вместо верного ответа 3 2 1 4 5 мы получим сумму штрафных баллов 0,9, а итоговый балл 0,1. Ввод последовательности 1 2 3 5 4 даст сумму штрафных баллов 0,6 и итоговый балл 0,4. Во втором ответе и правильный вариант ответа, и наиболее отстоящий от него дистрактор расположены не на своих позициях, в то время как в первом ответе абсолютно неверный ответ занимает правильную позицию. Тем не менее итоговый балл второго ответа оказался выше. Чтобы повысить значимость верных позиций правильного ответа и самого ошибочного из дистракторов, выбор которого, как правило, не требует знания предмета и основывается на общей эрудиции или каком-либо очевидном отличии от прочих вариантов, следует повысить значимость неверного указания первой и последней позиций вводимых последовательностей. Наиболее простой способ – увеличение значений α_1 и α_5 .

Положим, в нашем примере $\alpha_1 = \alpha_5 = 0,2$. Тогда штраф за первый ответ составит 1,1, а итоговый балл за выполнение ТЗУД, согласно правилу (3.20a), будет равен нулю. Штраф за второй ответ будет равен 0,9, а итоговый балл $B_1 = 0,1$.

Рассмотрим другой способ настройки алгоритма оценивания ТЗУД. Изменим показатель степени в правой части выражения (3.18)

для ужесточения штрафа за неверное указание крайних позиций в ответе. Значения $\delta_i = 1$ делают малозначимыми небольшие (в пределах одной позиции) отклонения номеров вариантов ответов вводимых последовательностей от эталонных. Это соответствует мнению разработчика о главенстве знания верного варианта над распределением позиций дистракторов, за исключением, возможно, последней, если соответствующий дистрактор содержит должное казаться очевидным абсурдное утверждение. Такая значимость первой и последней позиций в эталонных последовательностях повысится, если положить: $\delta_1 = \delta_5 = 0$. Изменение данного регулятора таким образом усиливает реакцию алгоритма начисления штрафных баллов за незнание единственного верного ответа. Для последовательности номеров вариантов ответа 1 3 4 2 5 в рассматриваемом примере мы получим итоговый штрафной балл 1,1 и $B_1 = 0$, а для последовательности 1 2 3 5 4 – итоговый штрафной балл 1,0 и $B_1 = 0$. Таким образом, количественные трактовки существенности допущенных ошибок в распределении позиций верного варианта ответа и дистракторов совпали.

Изменение значений β_i и γ_i обеспечивает настройку чувствительности алгоритма оценивания к ошибкам в трактовке их значимости составителем задания.

Более простой способ назначения штрафов заключается в табличном (матричном) задании их числовых значений P_i с учетом условий, отражающих значимость тех или иных ошибок в ее трактовке создателем теста. Простейшие примеры задания P_i :

$$P_1 = 0,5; P_N = 0,25; P_2 = P_3 = \dots = P_{N-1} = \frac{1}{4(N-1)}, \quad (3.22a)$$

$$P_1 = 0,4; P_N = 0,3; P_2 = P_3 = \dots = P_{N-1} = \frac{0,3}{N-1}, \quad (3.22b)$$

В (3.22a) и (3.22b) не учитывается, насколько далеко могут отстоять позиции вариантов ответа во введенной тестируемым и эталонной последовательностях. В итоге для ответа 1 3 4 2 5 получим итоговые баллы $B_1 = 0,25$ (3.22a) и $B_1 = 0,3$ (3.22b), а для ответа 1 2 3 5 4 – $B_1 = 0,08$ (3.22a) и $B_1 = 0,1$ (3.22b), т. е. качественное отличие от результатов применения алгоритмов начисления штрафов с помощью выражений (3.18), (3.20), (3.20a).

Детализация значений штрафов приводит к развертыванию строки вида (3.22a) в таблицу с N^2 значениями, т. е. существенному усложнению описания алгоритма вычисления баллов за решение ТЗУД.

В алгоритмах без начисления штрафных баллов табличный способ задания значений баллов A_i за частично верные решения, напро-

тив, является предпочтительным, так как начисление баллов идет только в случае совпадения цифр вводимой и эталонной последовательности, значения начисляемых баллов зависят от i , а разность $i - n$ не имеет значения.

Таблица 3.1

*Баллы для начисления за частично верные ответы
на тестовые задания с упорядочиванием дистракторов*

i	1	2	3	4	5
A_i	0,4	0,2	0,15	0,15	0,1

Значения A_i соответствуют трудности и значимости выбора верной позиции варианта ответа. В отличие от P_N , которой при начислении штрафов следует присваивать относительно большие значения, A_N принимает минимальное значение из всех A_i .

Применение табл. 3.1 к рассмотренным решениям примера 3.35 дает для последовательности 1 3 4 2 5 суммарный балл $B_1 = 0,1$, для последовательности 1 2 3 5 4 – $B_1 = 0,2$, что качественно согласуется с результатами применения алгоритма начисления штрафных баллов с использованием выражений (3.18), (3.20) и (3.20а).

Оценивание частично верных решений ТЗУД может основываться на соотнесении метрики – расстояния в метрическом пространстве между вводимой и эталонной последовательностями M или расстояния между точками (векторами) в N -мерном пространстве, координаты которых задаются номерами вариантов ответов (или связанной с этим расстоянием характеристикой), с некоторой нормой M_Δ (мерой), характеризующей допустимое отклонение (Δ – окрестность) полученного результата от желаемого [92, 93]. В этом случае:

$$B_T = B_0 \cdot T = \frac{M_\Delta}{M} \cdot T. \quad (3.23)$$

Величина M может рассчитываться как расстояние (метрика):

$$M = \sqrt{(n_1 - i_1)^2 + (n_2 - i_2)^2 + \dots + (n_N - i_N)^2},$$

или, что удобнее, как квадрат расстояния:

$$M = (n_1 - i_1)^2 + (n_2 - i_2)^2 + \dots + (n_N - i_N)^2. \quad (3.24)$$

Тогда допустимое отклонение, гарантирующее удовлетворительную оценку:

$$M_\Delta = N \cdot \Delta^2, \quad (3.25)$$

где Δ – задаваемый (регулируемый) параметр.

В табл. 3.2 представлены значения B_0 , рассчитанные для примера 3.35 и вводимых последовательностей 1 3 4 2 5 и 1 2 3 5 4 при трех различных значениях Δ .

Таблица 3.2

Δ	1 3 4 2 5	1 2 3 5 4
1	0,357	0,833
0,7	0,175	0,408
0,5	0,089	0,208

За счет относительно большой разности $n_3 - i_3$ величина M для первой последовательности оказывается больше, т. е. согласно табл. 3.2 ответ 1 3 4 2 5 отстоит дальше от истинного, чем ответ 1 2 3 5 4. Варьирование значением Δ позволяет регулировать масштаб оценочной шкалы.

Для усиления значимости неверного указания ключевых позиций, например с номерами 1 и N , можно применять масштабированные характеристики вида:

$$M = [\sigma_1 \cdot (n_1 - i_1)]^2 + (n_2 - i_2)^2 + \dots + [\sigma_N \cdot (n_N - i_N)]^2. \quad (3.24a)$$

В табл. 3.3 приведены значения B_0 , рассчитанные для рассматриваемых примеров с применением выражения (3.24a) при $\sigma_1 = \sigma_5 = 2$ и $\sigma_1 = 2, \sigma_5 = 2,5$.

Таблица 3.3

Δ	$\sigma_1 = \sigma_5 = 2$		$\sigma_1 = 2; \sigma_5 = 2,5$	
	1 3 4 2 5	1 2 3 5 4	1 3 4 2 5	1 2 3 5 4
1	0,167	0,200	0,167	0,182
0,7	0,082	0,098	0,082	0,089

Результаты расчетов демонстрируют возможность использования масштабных множителей σ_1 и σ_N в качестве регуляторов процедур начисления оценочных баллов за решение ТЗУД с возможным дальнейшим переводом их в оценки по Т-балльным шкалам путем масштабирования или применения критериев перевода, оперирующих числовыми пороговыми значениями или диапазонами.

Выбор регулировочных параметров в формулах (3.18), (3.19), (3.20), (3.20a), (3.22a), (3.22б) требует определенных ограничений и дополнительных условий во избежание ошибок оценивания (чрезмерного завышения начисляемого балла, несоразмерности начисляемых

и штрафных баллов и др.) при формальном соблюдении процедур. В то же время матричное представление начисляемых и штрафных баллов, получаемых методом подбора, при всех возможностях наглядного использования доступного диапазона числовых значений зачастую не поддается простому аналитическому описанию, что обуславливает формулирование дополнительных правил, устанавливающих взаимные соотношения между различными числовыми значениями параметров, которыми оперирует разработчик КТОЗУО [90].

Для формулирования и уточнения принципов оценивания нами были сформулированы вопросы для двух групп экспертов, представленных преподавателями вузов и учителями-предметниками старших классов средних школ, широко применяющих в своей практике дидактические тесты и знакомых с возможностями современных программных средств создания и реализации в образовательном процессе КТОЗУО. Вопросы касались общих подходов к оцениванию знаний с помощью ТЗУД и формулировались таким образом, чтобы формой ответа было однозначное «да» или «нет».

Итогом экспертного опроса стали следующие принципы разработки ТЗУД и оценочной шкалы и алгоритма начисления баллов, поддержанные более чем двумя третями экспертов в каждой группе [95]:

тестовые задания с упорядочиванием дистракторов позволяют оценивать не только формальное знание отдельных положений материала, но и его понимание;

тестовые задания с упорядочиванием дистракторов позволяют проверять системные знания в межпредметных областях;

тестовые задания с упорядочиванием дистракторов обладают потенциально высокой устойчивостью к применению различного рода недобросовестных приемов получения положительной оценки;

тестовые задания с упорядочиванием дистракторов позволяют адекватно оценивать реальные знания испытуемых при условии методически грамотного составления задания с учетом дидактических целей теста;

тестовые задания с упорядочиванием дистракторов позволяют формировать гибкие оценочные шкалы, адаптированные к задачам контроля знаний;

формулировка одного из дистракторов должна быть неправдоподобной или выходящей из общего ряда понятий;

формулировки дистракторов должны исключать неоднозначность их сопоставления на предмет близости к правильному варианту ответа;

формулировки вариантов ответа должны быть предельно лаконичны;

минимальная оценка за решение ТЗ должна составлять 0 баллов; знание верного ответа (первой позиции последовательности) не исключает получение 0 баллов за решение ТЗ;

знание верного ответа (первой позиции последовательности) должно обеспечивать положительную оценку, за исключением сочетания с размещением явно неверного (абсурдного) ответа выше предпоследней позиции;

общий положительный балл за решение ТЗ не может быть начислен при неправильном указании первой позиции последовательности;

начисление штрафов должно относиться ко всем неправильно указанным позициям последовательности;

наибольшие штрафы должны начисляться за несовпадение первой и (или) последней позиций ответа с эталонной последовательностью;

величина штрафа должна учитывать степень несовпадения каждой позиции в ответе с ее позицией в эталонной последовательности;

увеличение количества дистракторов должно сопровождаться повышением балльной оценки за правильное указание их позиции в последовательности;

балльная оценка абсолютно верного ответа ТЗ с меньшим количеством дистракторов должна быть больше балльной оценки частично верного ответа с большим количеством дистракторов.

На основании данных принципов нами были сформулированы следующие правила начисления баллов и штрафов [95]:

1) балл за правильное указание первой позиции последовательности не должен быть меньше суммы баллов за правильное указание всех остальных позиций последовательности за вычетом минимального начисляемого балла за правильное указание любой из позиций со второй по предпоследнюю;

2) балл за правильное указание последней позиции не может превышать любой из других баллов (меньше или равен наименьшему из баллов, установленных за правильное указание позиций последовательности со второй по предпоследнюю);

3) значение штрафа за неверное указание последней позиции последовательности должно быть меньше суммы баллов за правильное указание позиций, за исключением последней и предпоследней и за вычетом штрафа за единичное отклонение предпоследней позиции, на величину, не превышающую минимальный штраф за неверное указание любой позиции последовательности.

4) чем ближе к середине последовательности находится ее элемент, тем меньше максимально возможный штраф за неверное размещение этого элемента в последовательности;

5) величина штрафа за неверное указание каждой позиции, кроме первой и последней, должна быть пропорциональна абсолютному значению разности между номером этой позиции в эталонной последовательности и ее номером в последовательности ответа на ТЗ;

6) выбор коэффициента пропорциональности при реализации правила 5 должен обеспечивать выполнение правила 4.

7) балл за верное указание последней позиции не зависит от количества дистракторов;

8) балл за указание каждой верной позиции, кроме последней (следовательно, максимальная оценка за ТЗ), находится в прямой зависимости от количества дистракторов;

9) значение минимального штрафа за неверно указанную первую позицию последовательности равно сумме баллов за верное указание всех остальных позиций последовательности за вычетом минимальной суммы начисленного и штрафного баллов за неверное указание любой из позиций со второй по предпоследнюю.

Правила 7 и 8 содержат в себе условие нормировки начисляемых баллов. Их формулирование обусловлено тем, что отсутствие нормировки приведет к невозможности сопоставления оценок за задания различной трудности, которая, исключая содержательный аспект, определяется количеством дистракторов. Ориентируясь на данное условие, разработчик теста из ТЗУД может для удобства сопоставления ответов и упрощения алгоритма оценивания КТОЗУО в целом компенсировать содержательную трудность конкретного ТЗ уменьшением количества дистракторов и наоборот.

В свою очередь, нормирование на максимальную оценку (за верное указание первой позиции последовательности) повлечет за собой уменьшение начисленных баллов за правильное указание позиции дистракторов при увеличении их количества, т. е. при процедурном усложнении решения ТЗ для испытуемого, сопряженного с повышением вероятности ошибки первого рода.

Рассмотрим табл. 3.4, в которой представлены баллы для начисления за правильное указание позиций верного ответа и дистракторов в сформированной испытуемым последовательности при различных количествах дистракторов.

Таблица 3.4

№ позиции	1	2	3	4	5	6
Начисляемый балл	1,35	0,4	0,4	0,4	0,4	0,15
	0,9	0,35	0,35	0,35	0,2	
	0,7	0,3	0,3	0,1		

Значения начисляемых баллов удовлетворяют правилам 1, 2, 7, 8. Если верный ответ и дистракторы расставлены в правильном порядке, то для ТЗ с четырьмя вариантами ответа (один верный ответ и три дистрактора) суммарный балл составит 1,4, для ТЗ с пятью вариантами ответа – 2,15, для ТЗ с шестью вариантами ответа – 3,1. Очевидно, прямое суммирование дает чрезмерное повышение балла за решение ТЗУД по мере увеличения количества дистракторов. Чтобы уменьшить зависимость максимально набранного балла B_{max} за решение ТЗУД от количества дистракторов, нами было апробировано три варианта расчета B_{max} :

вариант А – расчет утроенного среднего значения:

$$B_{max} = \frac{3}{N} \cdot \sum_{k=1}^N B_k, \quad (3.26)$$

где B_k – балл за верное указание k -й из общего количества N позиций;

вариант Б – извлечение квадратного корня из суммы баллов:

$$B_{max} = \sqrt{\sum_{k=1}^N B_k}; \quad (3.27)$$

вариант В – извлечение квадратного корня из удвоенного среднего значения:

$$B_{max} = \sqrt{\frac{2}{N} \cdot \sum_{k=1}^N B_k}. \quad (3.28)$$

В табл. 3.5 приведены рассчитанные значения B_{max} .

Таблица 3.5

Количество дистракторов, N-1	Вариант А	Вариант Б	Вариант В	$\sum_{k=1}^N B_k$
3	1,05	1,18	0,84	1,4
4	1,29	1,47	0,93	2,15
5	1,55	1,76	1,02	3,1
Отношение max/min	1,48	1,49	1,21	2,21

Вариантов регулирования зависимости отношений $B_{\max}$ при разных количествах дистракторов может быть сколь угодно много. Разработчик теста, подбирая, например, степень корня из $B_{\max}$ (или $B_{\max}/N$) или основание логарифма $B_{\max}$ (или $B_{\max}/N$) и линейные коэффициенты, может получить практически любой желаемый результат, исходя из собственных трактовок влияния количества дистракторов на трудность вопроса и ее учете в максимальном начисляемом балле за абсолютно правильный ответ.

В табл. 3.6 приведены значения штрафов за неверное указание позиций верного ответа и дистракторов, удовлетворяющие правилам 3, 4, 5, 6, 9. Здесь Δn – абсолютная разность между номером позиции в эталонной последовательности и ответе испытуемого.

Таблица 3.6

№ позиции	1	2	3	4	5	6
Начисляемый штраф	1,25	$0,15 \cdot \Delta n$	$0,15 \cdot \Delta n$	$0,15 \cdot \Delta n$	$0,15 \cdot \Delta n$	2,3
	0,75	$0,15 \cdot \Delta n$	$0,15 \cdot \Delta n$	$0,15 \cdot \Delta n$	1,35	
	0,3	$0,1 \cdot \Delta n$	$0,1 \cdot \Delta n$	0,8		

Если в последовательностях из четырех, пяти и шести цифр все позиции указаны верно, за исключением последней и предпоследней, балл за решение ТЗ составит 0,1. Если не на своих местах находятся лишь первая и вторая цифра последовательностей из четырех или пяти цифр, то за решение ТЗ начисляется 0 баллов, из шести цифр – -0,05 балла (т. е. итоговый результат – 0 баллов). При расположении не на своих местах второй и предпоследней цифры в последовательности из четырех позиций получим 0,6 балла за решение ТЗ, для последовательности из пяти цифр в аналогичном случае балл за решение ТЗ составит 0,85, из шести цифр – 1,4. Если первая и последняя позиции определены верно, а остальные неверно, то минимальный балл за решение ТЗУД с пятью дистракторами составит 0,3, максимальный – 0,9, для ТЗУД с четырьмя и тремя дистракторами в аналогичных ситуациях начисляемые максимальные и минимальные баллы совпадут и составят соответственно 0,6 и 0,5.

Аналогичная работа была проведена нами по формированию оценочных шкал и алгоритмов начисления баллов за частично верные решения ТЗГО [95].

Пример 3.37. Расположите ответы по мере убывания от абсолютно верного. Ширина полосы пропускания резонансного контура определяется:

- 1) его частотным коэффициентом передачи;
- 2) его добротностью;
- 3) его импедансом;
- 4) относительной величиной мощности, рассеиваемой на его активном сопротивлении;
- 5) его реактивным сопротивлением на резонансной частоте.

Решение в целях удобства проверки, как и для ТЗУД, целесообразно представлять цифровой последовательностью. Верный ответ: 1 2 4 3 5. Объяснение решения. Модуль частотного коэффициента передачи представляет собой амплитудно-частотную характеристику, по уровню 0,707 от резонансного максимума которой определяется ширина полосы пропускания контура. Добротность резонансного контура определяет относительную ширину полосы его пропускания, для указания граничных частот полосы пропускания необходимо значение резонансной частоты. Обратная величина отношения теряемой в контуре мощности к запасаемой мощности в контуре за период и дает значение добротности, однако механизмы потерь мощности не ограничиваются рассеиванием на активном сопротивлении контура (мощность может теряться, например, за счет паразитной индуктивной связи с другими контурами), к тому же из формулировки ответа № 4 неясно, об отношении к какой величине идет речь – входной мощности или запасаемой за период. Для того чтобы вариант № 3 оказался безусловно верным ответом, необходимо говорить не об импедансе как таковом, а о его частотной зависимости. Значение реактивного сопротивления на резонансной частоте контура является еще менее достаточным параметром для определения полосы пропускания, чем импеданс.

В [6] представлен алгоритм оценивания ТЗГО, основанный на начислении одинакового количества баллов за каждую верно указанную позицию варианта ответа и штрафов, пропорциональных абсолютной разности номеров позиции в ответе испытуемого и эталонной последовательности. По нашему мнению, недостатком такого способа подсчета является возможность получения сравнительно высокого балла при неверном указании первой позиции последовательности, что следует трактовать как невыполнение необходимого условия решения ТЗ по существу.

По итогам экспертного опроса сформулированы следующие принципы разработки ТЗГО, оценочной шкалы и алгоритма начисления баллов [95]:

тестовые задания с градацией ответов позволяют оценивать не только формальное знание отдельных положений материала, но и их глубокое понимание;

тестовые задания с градацией ответов позволяют проверять системные знания в предметной области с возможностью охвата содержания нескольких тем;

тестовые задания с градацией ответов обладают потенциально высокой устойчивостью к применению различного рода недобросовестных приемов получения положительной оценки;

тестовые задания с градацией ответов позволяют объективно оценивать реальные знания испытуемых при условии методически грамотного составления заданий;

тестовые задания с градацией ответов позволяют формировать вариативные оценочные шкалы, ориентированные на задачи контроля знаний;

количество вариантов ответа ТЗГО должно составлять от четырех до пяти;

формулировки вариантов ответа не должны допускать неоднозначных смысловых толкований;

формулировки вариантов ответов ТЗГО должны исключать неоднозначность трактовок их полноты и корректности по отношению к абсолютно верному ответу;

при многообразии формулировок абсолютно верного ответа в ТЗГО должна присутствовать только одна из них;

формулировка одного из вариантов ответа должна быть на грани частично верного и неверного ответа;

минимальная оценка за решение ТЗ должна составлять 0 баллов; знание верного ответа (первой позиции последовательности) исключает получение 0 баллов за решение ТЗ;

общий положительный балл за решение ТЗ не может быть начислен при неправильном указании первой позиции последовательности;

начисление штрафов должно относиться ко всем неправильно указанным позициям последовательности;

наибольший штраф должен начисляться за несовпадение первой позиции ответа с эталонной последовательностью;

штраф за неверное указание последней позиции последовательности не должен быть меньше штрафа за неверное указание любой другой позиции;

величина штрафа должна учитывать степень несовпадения каждой позиции в ответе с ее позицией в эталонной последовательности;

увеличение количества вариантов ответа должно сопровождаться повышением балльной оценки за правильное указание их позиции в последовательности;

балльная оценка абсолютно верного ответа ТЗ с четырьмя вариантами ответа должна быть больше балльной оценки частично верного ответа с пятью вариантами ответа.

Данные принципы реализованы в табл. 3.7 (начисляемые баллы за совпадение позиций в эталонной последовательности и ответе на ТЗ) и 3.8 (штрафные баллы за неверное указание позиции варианта ответа).

Таблица 3.7

№ позиции	1	2	3	4	5
Начисляемый балл	1,2	0,6	0,3	0,2	0,1
	1,0	0,5	0,25	0,15	

Таблица 3.8

№ позиции	1	2	3	4	5
Начисляемый штраф	$0,5 \cdot \Delta n$	$0,1 \cdot \Delta n$	$0,1 \cdot \Delta n$	$0,1 \cdot \Delta n$	$0,2 \cdot \Delta n$
	$0,3 \cdot \Delta n$	$0,2 \cdot \Delta n$	$0,1 \cdot \Delta n$	$0,2 \cdot \Delta n$	

Итоговые суммы баллов могут подвергаться нормированию в целях реализации единой для всех видов ТЗ оценочной шкалы в пределах одного теста и учета трудности ТЗ. В частности, за верное решение ТЗУД или ТЗГО может начисляться больше баллов, чем за верное решение ТЗ с выбором единственно верного ответа при одинаковом количестве вариантов верного ответа в ТЗ, и т. п. Варьирование условиями нормирования суммарных баллов за решение ТЗ, значениями начисляемых баллов и штрафов позволяет разработчику КТОЗУО создавать оценочные шкалы в соответствии с его пониманием приоритетов оцениваемого содержания обучения, собственными трактовками оценивания частично верных ответов в сравнении с абсолютно верными решениями.

3.4. Практические рекомендации по обеспечению достоверности оценивания учебных достижений обучающихся компьютерными тестами

Помимо количества и структуры ТЗ, на вероятность ошибок первого и второго рода могут оказывать влияние многие факторы, которые необходимо учитывать разработчикам КТОЗУО. Основываясь на сформулированных классиком отечественной педагогической тестологии В.С. Аванесовым принципах (соответствия содержания теста целям тестирования, значимости, научной достоверности, соответствия уровню современного состояния науки, полноты тестируемых знаний, вариативности содержания, возрастающей трудности) [31, 35, 52, 53], приведем некоторые рекомендации, основанные на практическом опыте применения ТОЗУО авторами монографии.

1. Принцип синхронизации (учет степени актуальности проверяемых знаний и умений). Содержание и уровень трудности ТЗ должны учитывать фактор частичного забывания обучающимися содержания изученного материала [58, 59, 99, 100]. Фактор частичного забывания влияет на оценки тестируемых на контрольно-проверочных мероприятиях мониторинга качества реализации образовательных программ, проводимого образовательной организацией, внешнего аудита качества образования, педагогических экспериментов и др. Ошибки оценивания первого рода, порождающие неверные аналитические выводы об уровне освоения отдельных компетенций и образовательных программ в целом, могут быть обусловлены одинаковыми уровнями трудности и объемами проверяемого содержания для ДЕСО, изучение которых отстоит от момента проверки на разные, отличающиеся в разы временные интервалы. Следовательно, фактор забывания может быть учтен не только в содержании тестов и алгоритмах начисления баллов, но и в алгоритмах анализа результатов тестирования. В меньшей степени данный фактор влияет на результаты контроля остаточных знаний обучающихся [30, 98, 101–103], так как сроки проведения такого контроля обычно нивелируют относительную разницу упомянутых выше временных интервалов «момент окончания изучения – момент контроля», определяющую относительные объемы забывания изученного материала.

При конструировании КТОЗУО подлежит учету и восстановление или восполнение знания обучающимися отдельных ДЕСО, обусловленное повторным обращением к ним в содержании, изучаемом позднее.

2. Принцип однородности заданий. К дополнительным ошибкам первого рода приводит чрезмерное многообразие ТЗ, из которых составлен тест (например, наличие в нем ТЗ открытого типа или ТЗМВ наряду с ТЗВО). Несмотря на наличие пояснений (инструкций) или внешние отличия структуры ТЗ, тестируемым требуется время на «переключение» мыслительных стереотипов. В частности, возрастает вероятность ошибок из-за невнимательности, несвоевременного прочтения или недопонимания инструкций. Дополнительным условием правильного и быстрого восприятия ТЗ тестируемым является стилевая однородность формулировок заданий и вариантов ответов.

Рассмотрим пример стилевой неоднородности формулировок ответов.

Пример 3.38. Электрический ток в полупроводниках возникает:

- а) на границе p-n-перехода;
- б) при уменьшении ширины запрещенной зоны;
- г) если полупроводник – собственный;
- д) при наличии электронно-дырочных пар;
- е) под воздействием освещения.

При необходимости использования ТЗ различных видов следует делить КТОЗУО на блоки, в каждом из которых представлены ТЗ одного вида, или разбивать процедуру тестирования на этапы, каждый из которых представляет собой субтест, составленный из однородных ТЗ.

3. Принцип пирамиды «трудность – количество тестируемых». Совершенно не обязательно использовать большие ресурсы КТОЗУО на проведение мероприятий КУДО, разных по уровню сложности задач контроля. Уровень сложности задач КУДО во многом определяется уровнем мотивации и образовательных амбиций испытуемых. Для оценки знаний обучающегося, на протяжении всего периода изучения дисциплины балансирующего между оценками «неудовлетворительно» и «удовлетворительно», нет смысла применять тесты с большой дифференцирующей способностью, содержащие ТЗ различных уровней трудности. Трудность ТЗ, отобранных для формирования теста, можно представить четырьмя уровнями: простые, средней трудности, повышенной трудности, трудные. При этом для слабоуспевающего обучающегося достаточно включения в тест ТЗ первых двух уровней, ответы на ТЗ повышенной трудности и трудные он будет угадывать, что не скажется на его оценке (в особенности при адаптивном алгоритме оценивания), но может привести к снижению надежности теста, когда испытуемый по окончании КУДО перескажет запомненные им ТЗ и варианты ответа на них другим обучаю-

щимся. Принципы возрастающей трудности и соответствия содержания теста целям тестирования, сформулированные В.С. Аванесовым [52, 53], мы трансформировали в принцип трудности, адекватной целям контроля. При текущем контроле успеваемости, предварительной и промежуточной аттестации процедуры тестирования могут быть оптимально организованы по принципу пирамиды: тесты относительно низкого уровня трудности применяются к оценке знаний всех обучающихся, более трудные тесты – к оценке знаний тех из них, кто показал высокие результаты на предыдущем уровне. Данный подход позволяет модернизировать базы ТЗ частично, в основном в отношении сравнительно простых заданий, а также сокращать общее время тестирования и последующей перезачеменовки.

4. Принцип исключения тривиально правильных ответов. Выбор тестируемым правильного ответа среди предлагаемых вариантов нередко основан на зазубривании определений или подсказках и оказывается успешным даже при хорошем подборе дистракторов – испытуемый просто не вчитывается в дистракторы, не вникает в их смысл, сконцентрировавшись на поиске известной ему комбинации слов. Далеко не всегда удастся сформулировать ТЗ таким образом, чтобы оно не подразумевало тривиальный выбор правильного ответа, являющегося стандартизированным определением какого-либо понятия, и наоборот. Тем не менее необходимо стремиться к исключению такого рода ситуаций и формулированию ТЗ так, чтобы к узнаванию набора знакомых слов тестируемый в ходе мыслительного процесса добавлял поиск знаний из смежных тем или предметных областей, логику, общую и профессиональную эрудицию.

Пример 3.39. Электрический ток – это:

- а) *упорядоченное движение заряженных частиц;*
- б) движение электронов от катода к аноду под действием электрического поля;
- в) совокупность диффузионного и дрейфового механизмов переноса зарядов;
- г) колебания узлов кристаллических решеток вследствие изменения направления напряженности внешнего электрического поля;
- д) испускание электронов с поверхности вещества.

Пример 3.40. Упорядоченное движение заряженных частиц под действием электрического поля называется:

- а) *электрическим током;*
- б) электронным дрейфом;
- в) термоэлектронной эмиссией;

- г) массопереносом вещества;
- д) гальваническим эффектом.

Получить ответ на нетривиально сформулированный вопрос, набрав в поисковом окне интернет-браузера комбинацию слов, в отличие от приведенных выше примеров, удастся не всегда, а время попыток поиска в Интернете или получения иных подсказок со стороны ограничено, поэтому вероятность возникновения ошибки второго рода в ответе на ТЗ с нетривиальной формулировкой задания существенно уменьшается, а обучающийся, поставленный в условия нетривиальных формулировок ТЗ и ответов на них (пример 3.41) вынужден будет считаться с необходимостью более глубокого погружения в изучаемый материал.

Пример 3.41. Электрический ток может протекать без наличия:

- а) внешнего электрического поля;
- б) свободных носителей заряда;
- в) замкнутой электрической цепи;
- г) *металлического проводника*;
- д) постоянного знака градиента потенциала электрического поля.

5. Принцип иерархии дистракторов. Дистракторы, применяемые в ТЗВО или ТЗМВ, не должны быть в одинаковой степени похожи на правильный ответ, их «похожесть» на правильный ответ следует дифференцировать. Пример неудачного подбора дистракторов:

Пример 3.42. Электрический ток – это:

- а) *упорядоченное движение заряженных частиц*;
- б) движение зарядов под действием электрического поля;
- в) упорядоченное движение заряженных частиц в проводнике;
- г) движение заряженных частиц от одного электрода к другому;
- д) движение зарядов в замкнутой электрической цепи.

Помимо внешнего, терминологического сходства, ни один из дистракторов данного ТЗ не содержит полностью неверный ответ – все они являются частично правильными. Поэтому, анализируя выбор испытуемым ответа, сделанного осознанно или наугад, преподаватель при желании не сможет определить степень незнания предмета, что, наряду с собственно оцениванием, является одной из целей КУДО. В этом отношении ТЗ из примера 3.39 выглядит намного предпочтительнее. Выбор испытуемым варианта «г» из примера 3.39 свидетельствовал бы о полном непонимании им сущности и механизмов проявления электрического тока; выбор варианта «в» – скорее, о попытке угадать ответ, основываясь на «научнообразии» предлагаемого варианта, или о наличии обрывочных несистемных знаний; варианта «д» –

о серьезных пробелах в понятийной базе дисциплины и несформированности предметного кругозора, но в то же время о наличии общих представлений о природе электрического тока; варианта «б» – о правильных в целом представлениях о природе и проявлении электрического тока, не сформированных в реальное знание предмета, а также об игнорировании тестируемым лекций и рекомендованных преподавателем учебников.

6. Принцип соответствия времени решения виду и трудности задания. Разработчики КТОЗУО нередко пренебрегают дифференцированием времени ответа на ТЗ в зависимости от содержательной трудности и особенно структуры последнего. В то же время вполне достоверным представляется утверждение (в большей степени гипотетическое) о том, что на решение ТЗМВ требуется больше времени, чем на решение ТЗВО, при условии одинаковой содержательной трудности обоих ТЗ. Практика применения ТЗ различных видов свидетельствует о том, что решение ТЗУП требует как минимум столько же времени, сколько и решение ТЗМВ, а ТЗУД и ТЗГО – как правило, больше.

Задание времени ответа на ТЗ различных видов в КТОЗУО может быть определено на основе эмпирических данных или экспертных оценок.

Важно минимизировать время ответа на ТЗ с простой структурой (ТЗАО, ТЗ с дополнением, ТЗВО), ориентированного на проверку базовых или элементарных знаний. Автоматизм решения таких ТЗ косвенно свидетельствует о системности и высоком уровне знаний испытуемого.

7. Принцип проверки знаний, а не способностей. Формулировка ТЗ и дистракторов (парных соответствий, верных ответов, элементов последовательности), лимит времени решения должны быть ориентированы на оперативную проверку конкретного знания, а не способности испытуемого прийти к верному ответу путем умозаключений, синтеза знания и интуиции. В то же время во избежание ошибок первого рода, формулировки ТЗ и дистракторов ТЗВО, ТЗМВ и ТЗУД не должны быть громоздкими, изнуряющими внимание и память тестируемых, при этом незначительно отличаясь внешне или содержательно. Так, иногда каждый из дистракторов ТЗ на знание законодательных и иных нормативных правовых актов является объемным (три-четыре строки текста) и отличается от других одним словом или, по принципу фасетности, одним словом в каждом фасете. Испытуемый, не обладающий высокой концентрацией внимания, может допустить

ошибку в ответе на подобное ТЗ, даже будучи способным указать верный ответ в иных условиях.

8. Принцип минимизации использования фасетов в одном задании. Фасетная структура формулировки утверждения, истинность которого подлежит проверке, существенно упрощает составление ТЗ.

Пример 3.43. Согласно первому упоминанию в летописях считается, что Москва была основана Юрием Долгоруким (Ярославом Мудрым, Всеволодом Большое гнездо, Андреем Боголюбским) как город (село, крепость, столица, торжище) Ростово-Суздальского (Рязанского, Владимирского, Московского, Киевского) княжества в 1147 (988, 1208, 1237, 1480) г.

Однако, когда в ТЗ задействованы все фасеты, это не только увеличивает время его решения, но и затрудняет структурирование и анализ неверных решений по окончании тестирования. Для ТЗВО достаточно одного фасета, применение двух фасетов при формировании дистракторов может служить разработчику инструментом повышения трудности ТЗ.

9. Принцип редактируемости заданий. Выбор структуры ТЗ и формулирование вопросов (заданий) и ответов на них целесообразно осуществлять таким образом, чтобы их можно было видоизменять (редактировать вопросы (задания), менять или редактировать дистракторы, пары соответствий, верные ответы в ТЗГО) с минимальными трудозатратами. В этом заключается «ремонтпригодность» (по аналогии с техническими объектами) ТОЗУО, являющаяся одним из показателей их надежности. Регулярное редактирование ТЗ вместо их замены, во-первых, позволяет снижать трудоемкость процесса уменьшения вероятности ошибок второго рода, во-вторых, оставляя без изменения тестируемое содержание, не дает оснований пересмотра содержательной валидности теста.

10. Принцип контрольных «закладок». В своей практике разработки тестов мы используем ТЗ сверхтрудного уровня или с содержанием, выходящим за пределы изучаемого материала с соответствующим учетом в порогах и диапазонах оценок. Если процент решения таких ТЗ превышает вероятность случайного угадывания, это свидетельствует о том, что тест подвергся «взлому» и база ТЗ подлежит замене или редактированию. В разделе 2.2 механизм автоматизированной проверки теста на «взлом» реализован на основе сравнения процента решения простых и трудных заданий. Творческий подход к конструированию КТОЗУО обеспечивает широкое разнообразие подобного рода приемов – индикаторов повышения вероятности отказа теста.

ЗАКЛЮЧЕНИЕ

Разработка теста для контроля знаний и умений обучающихся, по сути, нацелена на определение области оптимальных решений в отношении количества тестовых заданий и уровней их трудности, распределения заданий по уровням трудности, времени решения отдельных заданий и теста в целом, отражения в заданиях приоритетов дидактических единиц проверяемого содержания обучения или степени содержательной валидности теста в целом и ряда других параметров оптимизации, обеспечивающих достижение требуемого целевого показателя – достоверности оценивания учебных достижений.

Ранее, анализируя надежность технических средств измерений и автоматических устройств обработки измерительной информации, таких как выбраковочные или сортировочные автоматы, тождественных по принципам функционирования компьютерным тестам для контроля знаний и умений обучающихся, мы определили надежность теста через вероятность его отказа, т. е. недостоверность оценивания учебных достижений, проявляющегося в ошибках первого и второго рода. Такой подход позволил устранить ряд методологических противоречий, например сведение надежности теста исключительно к воспроизводимости результатов его применения, что, в частности, приводит к снижению надежности теста при повышении его дифференцирующей способности. Он открывает перспективы применения методов теории надежности технических систем для получения априорных оценок надежности дидактических тестов, характеризующей как достижение требуемого уровня достоверности оценивания учебных достижений, количественно выражающегося в вероятности их верного оценивания, так и временной интервал или количество применений теста без его корректировки, в течение которого данный уровень будет актуален.

Прогнозирование и обеспечение требуемой надежности теста строится на основе оценок и повышения надежности входящих в его состав тестовых заданий в аспекте как их долговечности, так и минимизации вероятности ошибок оценивания, т. е. неправильного начисления баллов за верное, неверное или частично верное решение задания с учетом компенсации ошибок второго рода, порожденных угадыванием испытуемыми верных ответов при фактическом их незнании. Применение на этапе проектирования теста математических и экспертных моделей для получения априорных оценок его надежности дает возможность не только прогнозировать степень достоверности ре-

зультатов тестирования, но и обосновывать корректность и оптимальность выбора видов тестовых заданий, лимитов времени их выполнения, структуры теста и других регуляторов качества теста.

Развивая данный подход, мы предложили ряд способов оценивания частично верных ответов на тестовые задания усложненной структуры: с градацией верных ответов, с упорядочиванием дистракторов, на установление последовательности, что позволяет не только повысить дифференцирующую способность тестов, но и варьировать ее в зависимости от дидактических условий и целей контроля учебных достижений, формировать оценочные шкалы различной прецизионности. В качестве основной меры повышения надежности тестов рассмотрено включение в них тестовых заданий, наиболее устойчивых к ошибкам второго рода и деградации. Для подтверждения действенности данной меры результаты математического моделирования надежности тестовых заданий различных видов дополнены результатами их экспертизы на основе экспертной модели испытуемого. Применение тестовых заданий, обеспечивающих высокую надежность и одновременно компактность тестов, рассмотрено в аспекте целеполагания контроля учебных достижений, определяемого, в свою очередь, условиями реализации образовательных программ. Проанализированы перспективы использования таких заданий в алгоритмах адаптивного тестирования, для чего показана возможность регулирования трудности тестов с повышением количества уровней трудности с трех до шести применением в них различных видов заданий усложненной структуры без изменения их содержательной трудности, реализованная в нескольких оригинальных алгоритмах с перепроверкой результатов тестирования.

Рассмотренные в монографии меры повышения надежности компьютерных тестов для контроля учебных достижений обучающихся, наряду с примерами тестовых заданий, их корректировкой и моделированием их решений, практическими рекомендациями по обеспечению качества тестов, призваны активизировать творчество педагогов-практиков и результативность их работы в достижении высокой достоверности результатов оценивания знаний и умений обучающихся разрабатываемыми и модернизируемыми ими контрольно-измерительными материалами.

ЛИТЕРАТУРА

1. Ефремова Н.Ф. Тестовый контроль в образовании: учеб. пособие. – М.: Логос, 2005.
2. Беспалько В.П. Программированное обучение: дидактические основы. – М., 1970.
3. Талызина Н.Ф. Управление процессом усвоения знаний. – М.: МГУ, 1975.
4. Ким В.С. Некоторые источники погрешностей измерения уровня знаний в системе управления учебным процессом // Наука и школа. – 2010. – № 3. – С. 74–78.
5. Михайлычев Е.А. Дидактическая тестология. – М.: Народное образование, 2001.
6. Челышкова М.Б. Теория и практика конструирования педагогических тестов: учеб. пособие. – М.: Логос, 2002.
7. Ингенкамп К. Педагогическая диагностика. – М.: Педагогика, 1991.
8. Ким В.С. Тестирование учебных достижений. – Уссурийск: УГПИ, 2007.
9. Аванесов В.С. Применение тестовых форм в Rasch Measurement // Педагогические измерения. – 2005. – № 4.
10. Челышкова М.Б., Звонников В.И., Давыдова О.В. Оценивание компетенций в образовании: учеб. пособие. – М.: Изд. дом ГУУ, 2011.
11. Фишман И.С., Голуб Г.Б. Компетентностно-ориентированное образование: оценка результатов / под ред. Е.Я. Когана. – М.: ФИРО, 2015.
12. Соловова Н.В. Формирование и оценка компетенций: учеб. пособие. – Самара: Самарский университет, 2015.
13. Игнатьева И.Н. Компетентностно-ориентированные задания как средство оценки профессиональных компетенций студентов // Актуальные подходы к оцениванию компетенций студентов в условиях реализации ФГОС СПО: материалы Всерос. науч.-практ. конф. педагогических и руководящих работников / под науч. ред. Е.Е. Сасариной. – Ярославль: ЯГПУ, 2015.
14. Хулудеева И.В., Смирнова А.С., Баженов Р.И. и др. Оценочные средства для мониторинга сформированности компетенций в процессе естественно-математической подготовки в вузе: актуальные вопросы и опыт разработки / под общ. ред. Н.В. Эйрих. – Биробиджан: ПГУ им. Шолом-Алейхема, 2016. – 102 с.

15. Векслер В.А. Возникновение тестологии // Современные научные исследования и инновации. – 2015. – № 5. – Ч. 4. – URL: <http://web.snauka.ru/issues/2015/05/48321> (дата обращения: 26.11.2025).
16. РМГ 29-2013 ГСИ. Метрология. Основные термины и определения.
17. ГОСТ 16263-70. Государственная система обеспечения единства измерений.
18. РМГ 29-99 ГСИ. Метрология. Основные термины и определения.
19. Звонников В.И., Найденова Н.Н., Никифоров С.В., Челышкова М.Б. Шкалирование и выравнивание результатов педагогических измерений. – М.: Логос, 2003.
20. Балыхина Т.М. Словарь терминов и понятий тестологии. – М.: РУДН, 2000.
21. Татур А.О., Челышкова М.Б. Научно-методические проблемы создания системы тестирования в российском образовании // Развитие системы тестирования в России: тез. докл. Всерос. конф. – М., 1999. – Ч. 1.
22. Аванесов В.С. Форма тестовых заданий. – М.: Центр тестирования, 2005.
23. Опарина Н.М., Полина Г.Н., Файзулин Р.М., Шрамкова И.Г. Адаптивное тестирование: учеб.-метод. пособие. – Хабаровск: ДВГУПС, 2007.
24. Зайцева Л.В., Прокофьева И.О. Модели и методы адаптивного контроля знаний // Образовательные технологии и общество. – 2004. – Т. 7. – № 4. – С. 265–277.
25. Майоров А.Н. Теория и практика создания тестов для системы образования. – М.: Интеллект-центр, 2001.
26. Булгаков О.М., Старостенко И.Н., Хромых А.А., Дедикова А.О. Модели оценки качества тестов для контроля знаний. – Краснодар: Краснодарский университет МВД России, 2021. – 138 с.
27. Хлебников В.А., Михалева Т.Г. Отраслевой стандарт, педагогические тесты, термины и определения. – М.: Центр тестирования, 2001.
28. Булгаков О.М., Ладыга А.И., Рябошапка О.Н. Интерпретация результатов контроля остаточных знаний с применением элементов корреляционного анализа и математической статистики // Вестник Воронежского института ФСИИ России. – 2018. – № 2. – С. 33–37.

29. Булгаков О.М., Дедикова А.О. Математическая модель контроля безотказной работы теста для проверки знаний // Вестник Воронежского института МВД России. – 2018. – № 2. – С. 45–55.
30. Булгаков О.М., Ладыга А.И. Тестирование остаточных знаний обучающихся. – Краснодар: Краснодарский университет МВД России, 2021.
31. Аванесов В.С. Композиция тестовых заданий. – М.: АДЕПТ, 1998.
32. Сердюков В.И. Особенности адаптированного автоматизированного контроля знаний // Ученые записки ИИО РАО. – 2012. – № 40. – С. 62–73.
33. Челышкова М.Б. Теоретико-методологические и технологические основы адаптивного тестирования в образовании: дис. ... д-ра пед. наук. – М., 2001.
34. Татур А.О. Тесты в учебном процессе // Новые технологии в обучении и контроле знаний учащихся: материалы науч.-практ. конф. – М., 1999.
35. Аванесов В.С. Теория и практика педагогических измерений (материалы публикаций в открытых источниках и Интернет). – Екатеринбург: ЦТ и МКО УГТУ-УПИ, 2005.
36. Бочкарева Т.Н. Современные средства оценивания результатов обучения: учеб.-метод. пособие. – Махачкала: Апробация, 2019.
37. Мальцев А.В., Наймушина О.Э. Тестология в образовании: вчера, сегодня, завтра // Известия Уральского государственного университета. – 2008. – № 60. – С. 7–14.
38. Беспалько В.П. Методы оценки результатов тестирования учащихся // Вопросы тестирования в образовании. – 2007. – № 2. – С. 23–34.
39. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. – М.: Прометей, 2000.
40. Попов Д.И., Попова Е.Д. Экспертиза качества тестовых заданий: учеб. пособие. – М.: МГУП, 2008.
41. Мороз Л.С. Методы определения надежности и валидности тестов для контроля знаний // Труды БГТУ. Сер. VI: Физ.-мат. науки и информ. / под ред. И.М. Жарского. – Минск: БГТУ, 2010. – Вып. XVIII. – С. 176–179.
42. Моисеев В.Б., Усманов В.В., Таранцева К.Р., Пятирублевый Л.Г. Оценивание результатов тестирования на основе экспертно-аналитических методов // Открытое образование. – 2001. – № 3. – С. 32–36.

43. Рудинский И.Д., Литвинов К.А. Модель автоматизированного оценивания качества тестовых контрольно-измерительных материалов // Ученые записки ИИО РАО. – 2012. – № 40. – С. 74–90.
44. Энциклопедия психологических тестов. Личность. Мотивация. Потребность / под ред. А. Карелина. – М.: АСТ, 1997.
45. Анастаси А., Урбина С. Психологическое тестирование. – 7-е междунар. изд. – СПб.: Питер, 2001.
46. Челышкова М.Б., Звонников В.И., Татур А.О. Основные направления модернизации системы контроля и оценки качества учебных достижений учащихся // Квалиметрия человека и образования: методология и практика: тез. докл. X симп. – М.: ИЦПКПС, 2002. – Ч. 3.
47. Кларин М.В. Инновационные модели обучения в зарубежных педагогических поисках. – М.: Арена, 1994.
48. Летова Л.В. Точность моделирования латентных переменных с помощью модели Раша // Современные научные исследования и инновации. – 2014. – № 6. – Ч. 1. – URL: <http://web.snauka.ru/issues/2014/06/34399> (дата обращения: 24.11.2025).
49. Летова Л.В. Исследование качества теста как измерительного инструмента // Дистанционное и виртуальное обучение. – 2013. – № 11. – С. 116–125.
50. Маслак А.А., Моисеев С.И. Модель Раша оценки латентных переменных и ее свойства. – Воронеж: Научная книга, 2016.
51. Попов А.П. Критический анализ параметрических моделей Раша и Бирнбаума // Инновационные методы и средства оценки качества образования: материалы IV Всерос. науч.-метод. конф., Москва, 20–21 апр. 2006 г. – М.: МГУП, 2006. – С. 231–235.
52. Аванесов В.С. Научные проблемы тестового контроля знаний. – М.: ИЦПКПС, 1994.
53. Аванесов В.С. Основы педагогической теории измерений // Педагогические измерения. – 2004. – № 1. – С. 15–21.
54. Переверзев В.Ю. Технология разработки тестовых заданий: справ. руководство. – М.: Е-Медиа, 2005.
55. Булгаков О.М. Совершенствование структуры компьютерных тестов // Новые информационные технологии в процессе подготовки современного специалиста: межвуз. сб. науч. тр. – Липецк: ЛГПИ, 1999. – Вып. 2. – С. 13–20.
56. Булгаков О.М., Дедикова А.О. Тестирование учебных достижений: от проверки знаний к проверке понимания // Вестник Санкт-Петербургского университета МВД России. – 2020. – № 2(86). – С. 183–190.

57. Буймов А.Г. Закономерности поведения кривых забывания // Доклады ТУСУРа. – 2017. – Т. 20. – № 4. – С. 138–141.

58. Майер Р.В. Имитационное моделирование изучения студентами вузовского курса, учитывающее психологические закономерности усвоения и забывания // Научно-методический электронный журнал «Концепт». – 2015. – № 12. – С. 116–120. – URL: <http://e-koncept.ru/2015/15430.htm>

59. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов: учеб. / пер. с англ. Н.Н. Найденовой, В.Н. Симкиной, М.Б. Челышковой; под общ. ред. В.И. Звонникова, М.Б. Челышковой. – М.: Логос, 2010.

60. Денисенко Л.Г. Тестовые материалы в условиях реализации ФГОС для учреждений профессионального образования (из опыта разработки). – Новосибирск: Новосибирский институт мониторинга и развития образования, 2014. – Ч. 1.

61. ГОСТ Р ИСО 9000-2015. Системы менеджмента качества. Основные положения и словарь. – М.: Стандартинформ, 2015.

62. Рапопорт И.А., Сельг Р., Соттер И. Тесты в обучении иностранным языкам в средней школе. – Таллин: Валгус, 1987.

63. Коккота В.А. Лингводидактическое тестирование: науч.-теор. пособие. – М.: Высшая школа, 1989.

64. Allen J.P.B., Davies A. The edinburgh course in applied linguistics: testing and experimental methods. – London: Oxford University Press, 1977.

65. Анастаси А. Психологическое тестирование: в 2 кн. / пер. с англ.; под ред. К.М. Гуревича, В.И. Лубовского – М.: Педагогика, 1982. – Кн. 1.

66. Смирнова Г.И. Разработка тезауруса педагогических измерений Г. Раша // Педагогические измерения. – 2005. – № 4. – С. 62–64.

67. Иванова Л.А. Оценка уровня сформированности общих и профессиональных компетенций с помощью современных педагогических приемов // Молодой ученый. – 2016. – № 2. – С. 799–804.

68. Пугачев В.П. Тесты, деловые игры, тренинги в управлении персоналом: учеб. для студентов вузов. – М.: Аспект Пресс, 2003.

69. Гмурман В.Е. Теория вероятностей и математическая статистика: учеб. для прикладного бакалавриата. – 12-е изд. – М.: Юрайт, 2017.

70. Аванесов В.С. Основные понятия педагогической тестологии // Научные проблемы тестового контроля знаний: тез. докл. участников школы-семинара. – М.: ИЦПКПС, 1994. – С. 105–108.

71. Деменчёнок О.Г. Математические основы Rasch Measurement // Педагогические измерения. – 2010. – № 1.
72. Даммер М.Д., Рогозин С.А., Шамаева Т.Н. Задания в тестовой форме как средство диагностики методической подготовки будущего учителя физики. – Челябинск: Центр научного сотрудничества, 2013.
73. Лопаткина Е.В. Современные средства оценивания результатов обучения: учеб. пособие. – Владимир: ВлГУ, 2012.
74. Маслак А.А. Теория и практика измерения латентных переменных в образовании. – М.: Юрайт, 2016.
75. Булгаков О.М., Дедикова А.О. О применимости методологического аппарата теории надежности к оценке качества тестов для проверки знаний // Вестник Воронежского института ФСИИ России. – 2017. – № 4. – С. 214–222.
76. Павловская О.О. Статические методы оценки надежности программного обеспечения // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. – 2009. – № 26(159). – С. 35–37.
77. Острейковский В.А. Теория надежности: учеб. для вузов. – 2-е изд., испр. – М.: Высшая школа, 2008.
78. Ямпурин Н.П., Баранова А.В. Основы надежности электронных средств: учеб. пособие для студентов вузов / под ред. Н.П. Ямпурин. – М.: Академия, 2010.
79. Бугров Я.С., Никольский С.М. Высшая математика: учеб. для вузов: в 3 т. / под ред. В.А. Садовниченко. – 6-е изд., стер. – М.: Дрофа, 2004.
80. Булгаков О.М., Старостенко И.Н., Хромых А.А., Дедикова А.О. Оценка надежности теста для проверки знаний, составленного из заданий с множественным выбором правильных вариантов ответа // Вестник Воронежского института ФСИИ России. – 2019. – № 4. – С. 62–69.
81. Булгаков О.М., Старостенко И.Н., Хромых А.А., Дедикова А.О. Моделирование надежности тестов с усложненной структурой тестовых заданий // Вестник Воронежского института ФСИИ России. – 2020. – № 2. – С. 62–70.
82. Беспалько В.П. Педагогика и прогрессивные технологии обучения. – М., 2013.
83. Булгаков О.М., Ладыга А.И., Черненко А.А. Особенности целеполагания контроля учебных достижений обучающихся и его ре-

ализации в разработке дидактических тестов // Вестник Краснодарского университета МВД России. – 2022. – № 2. – С. 87–93.

84. Звонников В.И. Современные средства оценивания результатов обучения: учеб. пособие. – М., 2008.

85. Рудинский И.Д. Принципы интеллектуального автоматизированного тестирования знаний // Информационные технологии в образовании (ИТО-2001): материалы XI Междунар. конф. – М.: МИФИ, 2001.

86. Дуплик С.В. Интеллектуальные обучающие и контролируемые системы // Информатика. Информационные технологии. Средства и системы. – 2000. – № 2. – С. 87–90.

87. Булгаков О.М., Гривенная Е.Н., Дедикова А.О. Сравнительный анализ надежности тестовых заданий методом экспертного моделирования решения // Вестник Краснодарского университета МВД России. – 2023. – № 1. – С. 129–136.

88. Летова Л.В. Объективность измерений латентных переменных // Дистанционное и виртуальное обучение. – 2014. – № 3. – С. 83–94.

89. Булгаков О.М., Дедикова А.О. Практические рекомендации по обеспечению достоверности оценивания учебных достижений обучающихся компьютерными тестами // Национальная ассоциация ученых. – 2021. – № 65. – С. 25–28.

90. Булгаков О.М., Гривенная Е.Н., Дедикова А.О. Алгоритмизация оценивания решения тестовых заданий с упорядочиванием дистракторов // Вестник Краснодарского университета МВД России. – 2023. – № 2. – С. 130–137.

91. Свет Я.М. Колумб. – М.: Молодая гвардия, 1973.

92. Скворцов М.А. Примеры метрических пространств. – М.: МЦИМО, 2002.

93. Канатников А.Н., Крищенко А.П. Линейная алгебра: учеб. для вузов / под ред. В.С. Зарубина, А.П. Крищенко. – М.: МГТУ им. Н.Э. Баумана, 2002.

94. Латова Н.В. Опыт управления качеством высшего образования при помощи электронной системы обратной связи // Высшее образование в России. – 2011. – № 1. – С. 102–109.

95. Булгаков О.М., Гривенная Е.Н., Дедикова А.О. Особенности оценивания знаний с применением тестовых заданий усложненной структуры // Вестник Краснодарского университета МВД России. – 2025. – № 2. – С. 118–127.

96. Булгаков О.М., Гривенная Е.Н., Дедикова А.О. Разработка тестов для контроля знаний как способ организации проектной

деятельности обучающихся // Вестник Краснодарского университета МВД России. – 2025. – № 3. – С. 119–127.

97. Булгаков О.М., Гривенная Е.Н. Трехуровневая система оценивания учебных достижений обучающихся при проведении предварительной аттестации // Вестник Краснодарского университета МВД России. – 2024. – № 4. – С. 124–130.

98. Булгаков О.М., Гривенная Е.Н. Формирование тестов для контроля остаточных знаний обучающихся из фонда оценочных материалов // Вестник Краснодарского университета МВД России. – 2023. – № 4. – С. 150–155.

99. Майер Р.В. Учет изменения прочности знаний при обучении: моделирование в электронных таблицах Excel // Современные научные исследования и инновации. – 2015. – № 1. – Ч. 3. – URL: <http://web.snauka.ru/issues/2015/01/45010> (дата обращения: 27.03.2019).

100. Строганов В.Ю., Макаренко Л.Ф., Ярцев М.И., Ягудаев Г.Г. Модели забывания учебной информации для системы подготовки персонала // Теория и практика применения информационных технологий в промышленности и на транспорте: материалы IX Междунар. заочной науч.-практ. конф. молодых ученых, Москва, 12 нояб. 2013 г. – URL: auts.esrae.ru

101. Гибадуллина Р. Н. Контроль остаточных знаний // Вестник Казанского государственного энергетического университета. – 2013. – № 4. – С. 113–115.

102. Ершиков С.М., Иванова И.В. Мониторинг уровня остаточных знаний студентов медицинского университета // Ярославский педагогический вестник. – 2017. – № 5. – С. 139–144.

103. Денисова О.П. Подготовка студентов к контролю остаточных знаний на основе обобщающего повторения // Вектор науки Тольяттинского государственного университета. – 2012. – № 4. – С. 86–88.

104. Бедерханова В.П. Педагогическое проектирование инновационной деятельности: учеб. пособие. – Краснодар, 2018.

105. Земсков Ю.П., Асмолова Е.В. Основы проектной деятельности: учеб. пособие. – 2-е изд., стер. – СПб.: Лань, 2020.

106. Горобец Л.В. «Метод проекта» как педагогическая технология // Вестник Адыгейского государственного университета. Сер. 3: Педагогика и психология. – 2017. – № 2. – С. 48–59.

107. Мальшакова И.Л., Сорокин А.А. Организация проектной деятельности обучающихся в учебном процессе // Педагогическое образование и наука. – 2020. – № 1. – С. 56–63.

108. Булгаков О.М., Гривенная Е.Н., Ладыга А.И. Внешняя экспертиза тестов для контроля остаточных знаний // Вестник Краснодарского университета МВД России. – 2021. – № 2. – С. 127–133.

109. Селевко Г.К. Энциклопедия образовательных технологий: в 2 т. – М.: Народное образование, 2005. – Т. 1.

110. Стариченко Б.Е. Информационно-технологическая модель обучения // Образование и наука. – 2013. – № 4(103) – С. 91–111.

111. Загвязинский В.И. Стратегические ориентиры развития отечественного образования и пути их реализации // Образование и наука. – 2012. – № 4. – С. 3–15.

ОГЛАВЛЕНИЕ

Список сокращений	3
Введение.....	4
1. Тест как инструмент оценивания учебных достижений....	9
1.1. Технологии тестирования учебных достижений в современном образовании.....	9
1.2. Основные подходы к проектированию тестов для оценивания и контроля учебных достижений обучающихся.....	23
1.3. Виды тестовых заданий.....	38
1.4. Реализация обучающего потенциала тестов и тестовых заданий.....	56
1.5. Особенности разработки и применения тестов для адаптивного оценивания знаний.....	60
2. Математические модели оценки надежности компьютерных тестов для проверки знаний и умений обучающихся.....	69
2.1. Модели априорной оценки качества теста и достоверности результатов тестирования учебных достижений.....	69
2.2. Общий подход к моделированию надежности теста для контроля знаний и умений обучающихся.....	87
2.3. Учет видов тестовых заданий при моделировании надежности теста для контроля знаний и умений обучающихся.....	94
2.4. Сравнительный анализ надежности тестовых заданий методом экспертного моделирования решения.....	98
3. Некоторые вопросы повышения надежности и достоверности тестирования учебных достижений.....	110
3.1. Реализация целеполагания контроля учебных достижений обучающихся в разработке дидактических тестов.....	110
3.2. Применение тестовых заданий с повышенной устойчивостью к возникновению ошибок второго рода.....	120
3.3. Совершенствование оценочных шкал и алгоритмов начисления баллов за решение тестовых заданий.....	133
3.4. Практические рекомендации по обеспечению достоверности оценивания учебных достижений обучающихся компьютерными тестами.....	161
Заключение.....	167
Литература.....	169

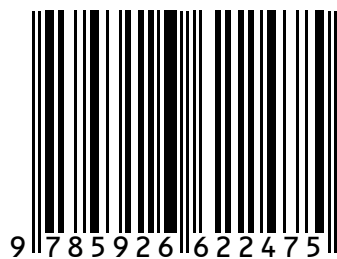
Научное издание

Булгаков Олег Митрофанович
Хромых Анна Алексеевна
Гривенная Елена Николаевна
Дедикова Анна Олеговна

**ОБЕСПЕЧЕНИЕ НАДЕЖНОСТИ ТЕСТОВ
ДЛЯ КОНТРОЛЯ ЗНАНИЙ**

Редактор *В. С. Ревина*
Компьютерная верстка *Г. А. Артемовой*

ISBN 978-5-9266-2247-5



Подписано в печать 31.03.2026. Формат 60x84 1/16.
Усл. печ. л. 10,5. Тираж 50 экз. Заказ 434.

Краснодарский университет МВД России.
350005, г. Краснодар, ул. Ярославская, 128.