Краснодарский университет МВД России

О. М. Булгаков
И. Н. Старостенко
А. А. Хромых
А. О. Дедикова

МОДЕЛИ ОЦЕНКИ КАЧЕСТВА ТЕСТОВ ДЛЯ КОНТРОЛЯ ЗНАНИЙ

Краснодар
2021

УДК 519.6
ББК 22.18
Б907

Одобрено
редакционно-издательским советом
Краснодарского университета
МВД России

Рецензенты:
Т. В. Мещерякова, доктор технических наук (Воронежский институт МВД России);
Н. А. Наумова, доктор технических наук (Кубанский государственный технологический университет);
А. Б. Сизоненко, доктор технических наук, доцент (Краснодарское высшее военное училище имени С. М. Штеменко);
А. Н. Прокопенко, кандидат технических наук, доцент (Белгородский юридический институт МВД России имени И. Д. Путилина);

Булгаков О. М.
Б907 Модели оценки качества тестов для контроля знаний / О. М. Булгаков, И. Н. Старостенко, А. А. Хромых, А. О. Дедикова. – Краснодар : Краснодарский университет МВД России, 2021. – 138 с.

ISBN 978-5-9266-1728-0

В монографии рассматриваются теоретические и практические проблемы тестирования учебных результатов. Раскрываются вопросы математического моделирования надежности компьютерного теста с различной структурой единичных тестовых заданий.

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

УДК 519.6
ББК 22.18

ISBN 978-5-9266-1728-0

Введение

Широкое применение дистанционных образовательных технологий во всех без исключения образовательных структурах как в России, так и за рубежом, обусловленное мерами профилактики COVID-19, актуализировало проблему объективного контроля и оценивания результатов учебной деятельности обучающихся и придало ей всеобщий характер. Применение для оценивания учебных достижений обучающихся онлайн-технологий, реализуемых, как правило, средствами видео-конференц-связи, выявило ряд проблем, порожденных как увеличением масштабов их использования в рамках каждой образовательной организации, так и общим невниманием научно-педагогической общественности к онлайн-дидактике в предшествующие годы. Проблемы применения онлайн-технологий для оценивания знаний и умений обучающихся можно условно разделить на две группы:

– *организационно-технические*, обусловленные ограниченной пропускной способностью каналов интернет-связи и производительностью серверов видео-конференц-связи, лимитирующие количество активных участников контрольно-оценочных мероприятий и приводящие к сбоям в передаче видеосигналов;

– *методические*, выражающиеся в стремлении преподавателей перенести в специфичный формат видео-конференц-связи типовые методики проверки знаний, например в ходе промежуточной аттестации, во время семинаров и практических занятий, основанные на непосредственном общении педагога с обучающимися, без их надлежащей адаптации к возможностям диалога в Интернете. Как правило, вследствие разного рода сбоев в трансляции сигнала время онлайн-опроса одного обучающегося в среднем в полтора-два раза больше, чем при общении «вживую». Кроме того, экзаменуемые нередко прерывают связь с экзаменатором, имитируя технический отказ видеоканала, а исключить несанкционированное использование ими различных источников информации как непосредственно при ответе, так и при подготовке к нему практически невозможно.

Таким образом, ключевым условием объективности оценивания результатов учебной деятельности обучающихся в дистанционном режиме становится оперативность контрольно-оценочных мероприятий, т. е. их проведение в такие сроки, которые, с одной стороны, исключили бы или существенно затруднили применение различного рода шпаргалок недобросовестными обучающимися, а с другой – снизили вероятность прерывания сеанса видео-конференц-связи.

Данное требование обуславливает применение компьютерных тестов для объективного оценивания учебных достижений обучающихся как в качестве самостоятельного инструмента, так и инструментальной основы предварительного, предшествующего онлайн-диалогу, контроля в многоэтапных методиках оценивания знаний и умений.

Не менее важной задачей оценивания учебных достижений обучающихся, обуславливающей повышенную востребованность в контрольно-проверочных мероприятиях компьютерных тестов, является необходимость достоверного определения уровня освоения (достижения) обучающимися компетенций, предусмотренных 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 выпускников представляет оценка уровня освоения компетенций обучающимися на завершающем этапе обучения, непосредственно перед Государственной итоговой аттестацией и в ее процессе. 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.

При обоснованном предпочтении компьютерных тестов на основе заданий закрытого типа, продиктованном их наилучшей встраиваемостью в методики оперативного оценивания и анализа

учебных достижений обучающихся, однозначностью трактовок ответов обучающихся и исключению влияния субъективного отношения педагога к обучающемуся на оценку его результата обучения, проблема обеспечения приемлемой надежности компьютерного тестирования в условиях расширенного применения дистанционных образовательных технологий существенно обостряется. Основная причина этого – уменьшение среднего времени применения тестов из-за повышения действенности их коллективного «взламывания» обучающимися в условиях отсутствия дистанционного контроля или его низкой эффективности. Так, первые тестируемые, имея возможность не контролируемого преподавателем копирования вопросов (путем скриншота или фотографирования экрана), отсылают вопросы теста и ответы, в которых они уверены, своим товарищам, те после своего тестирования дополняют базу вопросов (если компьютерный тест для оценки освоения ими некоторого содержания обучения генерируется выборкой из базы заданий) и ответов и т. д. Как показывает практика, через 3–4 цикла применения исходный компьютерный тест, составленный из заданий с выбором единственно верного ответа, оказывается «взломанным» на 50 % и более. Таким образом, при налаженном взаимодействии подвергаемых контролю обучающихся, в зависимости от количества заданий теста $N_{Зi}$ и отношения $N_{Зi}/N_З$, где $N_З$ – общее количество заданий в базе, из которой делается выборка $N_{Зi}$, при тестировании последовательно отдельных групп обучающихся, уже в третьей по счету группе средний балл оценки увеличивается примерно на 1, а порядка 20% правильных ответов становятся известными обучающимся уже из первой тестируемой группы. При последовательном тестировании отдельных обучающихся, начиная с шестого по счету тестируемого, оценка не опускается ниже «удовлетворительно», а средний балл любых пяти выбранных подряд тестируемых, начиная с десятого номера, как минимум на 1 выше, чем у первой пятерки. Применение более сложных тестовых заданий, например с множественным выбором верных ответом, установлением соответствий или правильных последовательностей, затрудняет и замедляет «взлом» теста, однако не обеспечивает его время безотказной работы, характеризующейся приемлемой достоверностью оценок знаний и умений обучающихся, достаточное для приме-

нения хотя бы для одной зачетно-экзаменационной сессии, проводимой исключительно на основе дистанционных технологий для контингента обучающихся численностью 50 и более человек. Данное обстоятельство вынуждает разработчиков тестов постоянно перерабатывать базы заданий, изменять процедуры тестирования, что повышает трудоемкость применения данного инструмента контроля успеваемости и тем самым нивелирует его преимущества относительно других форм оценивания знаний и умений обучающихся.

Таким образом, обеспечение надежности компьютерного теста для оценивания учебных достижений обучающихся, характеризующейся приемлемыми значениями среднего времени его применения или вероятности объективного (например, согласованного с другими формами контроля) оценивания знаний и умений, становится ключевым условием реализации дистанционной формы тестирования, как правило, не позволяющей преподавателю визуально контролировать аудиторию даже в онлайн-формате. Среди способов решения данной задачи наиболее очевидным представляется увеличение базы заданий с одновременным увеличением как $N_З$ и $N_{Зi}$ по отдельности, так и $N_{Зi}/N_З$. Данный способ оказывается действенным в сочетании с модифицированными методиками тестирования, включающими в себя, например, несколько этапов оценивания с корреляционным анализом их результатов или сопоставление результатов компьютерного тестирования с оценками других, независимых форм контроля (например, текущей успеваемости, контрольных работ и т. д.). Явным недостатком этого способа является увеличение времени контроля и формирования оценки, сводящее на нет преимущества компьютерного тестирования перед онлайн-собеседованием.

Организационно-технологические особенности тестирования учебных достижений в дистанционном формате, такие как невозможность критического анализа ответов, их обсуждения с преподавателем, апелляции и других форм диалога, обеспечивающих объективность оценок и ее осознание обучающимися, являются фактором снижения действенности и оперативности обратных связей в учебном процессе и тем самым повышают цену ошибки формирования или выбора комплекта контрольно-измерительных материалов как в содержательном, так и в алго-

ритмическом аспекте. Это, в свою очередь, актуализировало проблему априорной оценки качества компьютерных тестов, в общем случае характеризующегося степенью достоверности оценок учебных достижений. Решение данной проблемы, являющееся необходимым условием повышения надежности тестов, осложняется многообразием подходов к выбору и определению показателей, характеризующих качество теста, как правило, не реализованных в адекватных математических моделях.

Создание таких моделей необходимо для корректного анализа различных видовых тестовых заданий в целях выявления среди них обладающих относительно высокой потенциальной надежностью за счет их структуры и возможности оперативной (с минимальными трудозатратами) переработки, направленной на восстановление контрольных и дидактических функций теста, подвергшегося попытке «взлома». Применение потенциально высоконадежных тестовых заданий, по нашему мнению, обладает большей перспективой повышения достоверности оценок тестирования учебных достижений и регулирования сложности тестов по сравнению с увеличением объемов баз заданий и применением сложных эвристических и эмпирических критериев оценивания результатов выполнения тестов.

Для решения задачи повышения достоверности тестирования учебных достижений в монографии последовательно рассмотрены особенности современного этапа разработки и применения педагогических тестов (глава 1), проанализированы понятийная база и методология оценки их качества, предложен подход к моделированию оценок показателей качества тестов на основе теории надежности технических систем и ее математического аппарата (глава 2), проведен сравнительный анализ надежности тестов, составленных из наиболее распространенных видов тестовых заданий закрытого типа (глава 3), предложен новый вид тестового задания закрытого типа, отличающийся повышенной устойчивостью к «взлому» и высоким относительно других видов тестовых заданий дидактическим потенциалом, получены оценки надежности тестов, составленных из таких заданий, с различными алгоритмами подсчета баллов за правильные ответы и даны практические рекомендации по совершенствованию структуры и содержания тестов (глава 4).

1. Современные средства оценивания результатов обучения

1.1. Тест как инструмент оценивания учебных достижений

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 и др., вошедшая в отечественную образовательную практику с начала 90-х гг. прошлого столетия, стала одним из ведущих направлений модернизации контрольно-оценочных процедур, стандартизации требований на входе-выходе на разных уровнях обучения, обеспечения доступности профессионального образования, мониторинга освоения образовательных программ, оценки эффективности образовательных систем и их структурных элементов, повышения качества обучения [1]. За последние годы в нашей стране в этом направлении произошли большие изменения. Так, переход к единому государственному экзамену (ЕГЭ) обеспечил прозрачность для общества оценки результатов учебного процесса всей отечественной системы общего образования. 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. Проведение промежуточной и итоговой аттестации в формате компьютерного тестирования при реализации образовательных программ дополнительного профессионального образования обеспечило широкое внедрение дистанционных образовательных технологий в данный образовательный сегмент и позволило за счет сокращения временных и ресурсных затрат на обучение максимально удовлетворить потребности работников и их работодателей в повышении квалификации и профессиональной переподготовке.

Тестирование знаний и умений обучающихся, в том числе на основе компьютерных технологий, следует рассматривать как разновидность педагогических измерений, имеющих более широкое предназначение: например, наряду с количественными параметрами учебных достижений, выявление мотивации к обучению, степени соответствия условий обучения образовательным целям.

Большой вклад в развитие теории и практики педагогических измерений внесли многочисленные 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.

Многие специалисты в области оценивания учебных достижений отмечают, что одной из причин деградации образования является плохая организация системы наблюдения за результативностью учебного процесса, в результате чего, с одной стороны, наблюдается тотальный либерализм, а с другой – субъективизм и предвзятость.

При оценивании учебных достижений обучающихся многие специалисты ориентируются на тестирование их знаний и умений как наиболее объективную и независимую систему измерения, предоставляющую возможность массовой, оперативной, многомерной диагностики результатов учебно-познавательной деятельности.

Тесты как объективное диагностическое средство предоставляют информацию для анализа процессов в отдельных образовательных системах и сфере образования в целом, характеризующуюся такими качествами, как точность, полнота, достаточность, системность, оптимальность, обобщенность, оперативность и доступность.

Для понимания современных методов и технологий оценивания знаний и умений необходимо рассмотреть основные понятия и термины, связанные с тестологией, в порядке, отражающем последовательность действий от планирования и моделирования теста до его применения и обработки результатов.

Тестология (в переводе с англ. *test* – проба, греч. *logos* – знание) имеет несколько значений: с одной стороны – это наука

об испытании, с другой – наука о методах педагогических измерений и их интерпретации [2].

Проанализированные нами определения *педагогических измерений* выявили характерную в целом для данной отрасли квалиметрии системную проблему, порожденную несоответствием употребляемых терминов и понятий их аналогам в других отраслях знания. Такая терминологическая многозначность, а подчас 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 десятилетиями и веками и стандартизированных толкований терминов аналогичного написания и звучания.

Так, согласно рекомендациям РМГ 29-2013 [3], *измерение* – это процесс экспериментального получения одного или более значений величины, которые могут быть обоснованно приписаны величине. В рекомендациях предельно обобщены вводимые ранее различными стандартами [4, 5] 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 [6]: *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) выбор методов измерения и измерительных процедур (методик);

3) выбор (конструирование) измерительных инструментов и их интегрирование в измерительные процедуры;

4) выбор (конструирование) измерительной шкалы;

5) получение массива экспериментальных данных и их формализация, т. е. присвоение им первичных числовых значений;

6) отображение результатов измерения на измерительную шкалу по определенным правилам (алгоритмам);

7) обработка и интерпретация результатов измерения.

В теории педагогических измерений существует много определений понятия *педагогического теста*, заметно отличающихся друг от друга, но до сих пор нет ни одного самодостаточного, четкого и однозначного, признаваемого большинством педагогического сообщества.

Т.М. Балыхина определяет *педагогический тест* как комплекс заданий, измеряющих уровень учебных достижений, обученности, прогресс в учебной деятельности, эффективность учебного процесса [7]. Перечисленные объекты измерений очень разнородны в определении возможных измеряемых параметров. Очевидно, «уровень учебных достижений» и «прогресс в учебной деятельности» 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, не обеспечивающееся неким «комплексом заданий».

В определении, приведенном в [8], *педагогический тест* – это квалиметрически выверенная система тестовых заданий, методов их предъявления и оценивания результатов их выполнения, которая обеспечивает получение наиболее обоснованных характеристик объекта испытания. Данное определение опирается на неоднозначно трактуемые формулировки, такие как «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 по сути является тест.

В.С. Аванесов предлагает следующее определение: *педагогический тест* – система параллельных заданий возрастающей трудности, специфической формы, которая позволяет качествен-

но и эффективно измерить уровень и структуру подготовленности испытуемых [9]. В данном определении спорной является возрастающая трудность заданий, устанавливающая не столько дифференцирующую способность теста, сколько алгоритм тестирования «от простого к сложному», мягко говоря, не часто реализуемый в тестовых процедурах. 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 дидактической единицы содержания обучения (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 применение заданий открытого типа (например, используемых в разделах «b» и «c» тестов ЕГЭ) позволяет применить рассматриваемое определение к любой форме контроля знаний и умений, реализуемой письменно или на компьютере.

В своих рассуждениях мы будем исходить из того, что педагогическое измерение является более общим понятием по отношению к измерениям учебных достижений. Следовательно, педагогический тест является более общим понятием по отношению к тесту для оценивания учебных достижений.

Под *тестом для оценивания учебных достижений* (ТОУД) 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.

Согласно определению ТОУД включает в себя три компонента:

– задания, систематизированные по какому-либо принципу (например, отобранные в соответствии с ДЕСО, распределенные по уровням сложности, ориентированные на применение различных или однотипных форм тестовых заданий);

– процедуру перевода ответов испытуемых в баллы или иные оценочные показатели – оценивание в баллах или иных подлежащих дальнейшей обработке и анализу информативных показателях (цветах, символах, отождествляющих эмоции «смайликах» и др.) отдельных тестовых заданий в зависимости от их уровня сложности и типа;

– процедуру перевода полученных баллов или иных информативных показателей в итоговую оценочную шкалу, однозначно характеризующую уровень проверяемых учебных достижений.

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

Перевод ответов испытуемых в баллы для каждого тестового задания производится на основе сличения ответа с эталонным решением этого задания (эталоном), хранящемся 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 с помощью 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 [10].

По нашему мнению, реализация ТОУД на базе компьютерных технологий накладывает на структуру и содержание заданий, процедуры оценивания ряд специфических ограничений. Так, в компьютерных тестах ввиду ограниченных возможностей проверочных алгоритмов практически не используются прямые вопросы (задания открытого типа), подразумевающие развернутые ответы или ответы из нескольких слов, задания творческого характера, исключается употребление синонимов в ответах на прямые вопросы. Таким образом, компьютерное тестирование позволяет 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 ТОУД на базе компьютерных технологий определим *как компьютерный тест для оценивания знаний и умений обучающихся* (КТОЗУО).

В [11] *процедура тестирования* рассматривается как целенаправленное, одинаковое для всех испытуемых обследование, проводимое в строго контролируемых условиях и позволяющее объективно измерить изучаемые характеристики испытуемого и педагогического процесса. На наш взгляд, данному определению в большей степени соответствует термин «тестирование», в то время как процедуру следует трактовать как реализацию взаимообусловленной последовательности операций в заданных (контролируемых) условиях, т. е. понятие, приближенное к алгоритму. Спорным в приведенном определении является и требование одинаковости для всех испытуемых, исключающее индивидуальный подход к оцениванию знаний и умений обучающихся на основе априорных данных о степени успешности освоения ими учебного материала. По нашему мнению, придание процедуре тестирования отдельного терминологического смысла является избыточным.

Единичное (отдельное, обособленное) задание теста как структурный элемент базы заданий или ее выборки для конкретного теста в работах различных авторов имеет различное название, трактовку и признаки. Так, в [12, 13] для обозначения элемента базы заданий теста употребляется термин *педагогическое задание*. Нами были опрошены более двухсот человек, имеющих прямое отношение к педагогическому процессу в вузе (профессорско-преподавательский состав, адъюнкты очной и заочной адъюнктуры, обучающиеся старших курсов очной формы направлению «Психология и педагогика», специалисты управления учебно-методической работы и учебных отделов) на предмет понимания ими словосочетания «педагогическое задание». Все без исключения преподаватели составляли КТОЗУО при реализации программ высшего или (и) дополнительного профессионального образования. Около сорока процентов опрошенных посчитали педагогическим заданием поручение педагогу на выполнение какой-либо работы: разовой – проведения учебного заня-

тия, в том числе открытого или показательного, педагогического эксперимента, научно-практического исследования, разработки методического обеспечения преподавания учебной дисциплины, или на регулярной основе – выполнение учебной нагрузки на учебный год или семестр, проведение какого-либо вида занятий по учебной дисциплине в течение учебного года (семестра) и т. п. Немногим более трети опрошенных отождествили педагогическое задание с педагогической задачей: в широком смысле – с задачей педагога или образовательной организации по достижению целей образовательного процесса, в узком смысле – с планируемым мероприятием по реализации частных задач обучения. Оставшиеся (примерно четверть опрошенных) посчитали педагогическим заданием то, что задает преподаватель обучающимся для самостоятельной подготовки (реферат, доклад, домашняя работа) или решения в аудитории. Одиннадцать (менее пяти процентов) опрошенных в числе решаемых в аудитории заданий указали тест, в том числе компьютерный. Ни один из опрошенных не отождествил данный термин именно с отдельным заданием теста, что еще раз подтвердило избыточность терминов в отечественных научных исследованиях по вопросам тестирования учебных достижений, искусственность отдельных из них, субъективизм определений, порожденный собственным практическим опытом и индивидуальными трактовками оптимальности процедур разработки тестов, и отсутствие в целом системного научного подхода к формированию устойчивого понятийного аппарата в данной отрасли научного знания.

Ряд авторов наделяют *тестовые задания* такими характеристиками, как трудность, дифференцирующая способность, вариативность, локальная независимость, технологичность и эффективность, проверяемыми эмпирическим путем. Без такой апробации задания не могут быть тестовыми, а являются *претестовыми*. Требование известной трудности заданий является важнейшим системообразующим признаком тестового задания. На наш взгляд, такие параметры, как технологичность, вариативность и локальная независимость, могут быть определены априори, умозрительно – самим разработчиком теста или экспертами. Отчасти это относится и к трудности и дифференцирующей способности, а эффективность теста или отдельного задания может быть оце-

нена только по результатам завершенного исследования. Оценки эффективности, трудности и дифференцирующей способности, основанные на пробных тестированиях, чаще всего являются несостоятельными, что, например, укрепляет убежденность сторонников формирования тестов на основе двухпараметрической модели Раша [14–16] в том, что несоответствие модели практическим данным является проблемой практики, а не теории [17].

Объективности оценок теста способствует включение в него заданий разных видов.

Цели и задачи тестирования обусловливают *структуру теста,* которая определяет количество и название частей теста, соотнесенных с проверяемыми разделами содержания обучения, количество и последовательность тестовых заданий.

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

По новизне методических принципов разработки и дидактических условий реализации образовательного процесса, неотъемлемым элементом которого является оценивание учебных достижений обучающихся, тесты подразделяются на *традиционные* и *нетрадиционные*.

Традиционный тест реализует метод диагностики испытуемых, в котором они отвечают на идентичные задания, в одинаковое время, в одинаковых условиях и с одинаковой оценкой. К традиционным относятся гомогенные и гетерогенные тесты.

Гомогенные тесты предназначены для оценки учебных достижений испытуемых по одной учебной дисциплине или ее части. Требование дисциплинарной чистоты теста определяет содержание его заданий.

Гетерогенные тесты используются для оценки учебных достижений испытуемых по нескольким учебным дисциплинам. Гетерогенный тест может состоять из нескольких гомогенных субтестов [18]. Такие тесты применяются для оценивания сформированности у обучающихся сложных компетенций, комплексной оценки уровня подготовки выпускников образовательной организации и др. Для интерпретации результатов тестирования применяются усложненные алгоритмы формирования и агрегирования баллов.

К нетрадиционным тестам относятся интегративные, адаптивные, многоступенчатые и критериально-ориентированные.

Интегративные тесты включают в себя задания, отвечающие требованиям интегративного содержания, возрастающей трудности, нацеленные на обобщенную итоговую диагностику учебных достижений выпускников образовательной организации, например в ходе их Государственной итоговой аттестации [9]. Правильные ответы на задания интегративного теста требуют обобщенных и взаимосвязанных знаний двух и более учебных дисциплин. Преимущество интегративных тестов перед гетерогенными заключается в большей содержательной информативности каждого задания и в меньшем числе самих заданий.

В *адаптивном тесте* каждое последующее задание подбирается автоматически, специальным адаптивным алгоритмом, по результатам решения предыдущих заданий с учетом определенного заранее уровня их сложности. В начале тестирования испытуемый, по своему или алгоритмическому выбору, решает задания, с которыми он в состоянии справиться. Если испытуемый дает верный ответ, ему предлагается более сложное задание. При неправильном ответе следующее задание выбирается из более легких. При адаптивном тестировании обеспечивается индивидуализация набора заданий, контролируется не только правильность ответов, но и время, затраченное на решение, поддерживается высокий уровень мотивации к выполнению теста у менее подготовленных обучающихся за счет исключения сложных заданий. Считается, что адаптивные тесты, как и адаптивные автоматизированные алгоритмы измерения и контроля параметров технических объектов, обладают потенциально минимальной погрешностью измерения, применительно к тестам – оценивания уровня учебных достижений каждого испытуемого.

В зарубежной литературе выделяется несколько разновидностей адаптивного тестирования. При *пирамидальном тестировании* в отсутствие предварительных оценок всем испытуемым дается задание средней трудности и уже затем, в зависимости от ответа, для каждого испытуемого уровень трудности заданий повышается или понижается. В другой разновидности испытуемый сам выбирает уровень сложности заданий, а затем этот уровень постепенно приближается к реальному уровню знаний. Еще в одном

распространенном варианте реализации адаптивного тестирования испытуемые последовательно решают задания возрастающей сложности, а сложность и содержание заданий оставшейся части теста определяются исходя из результатов начального этапа [11].

По целям и подходам к интерпретации результатов, определяющим структуру базы заданий, процедуры тестирования и анализа результатов, различают тесты нормативно-ориентированные, критериально-ориентированные, содержательно-ориентированные.

Нормативно-ориентированный тест предназначен для распределения испытуемых по уровням подготовленности или порядковым местам. Все испытуемые за одинаковое время в одинаковых условиях отвечают на одинаковые задания. Их результаты оцениваются на одной и той же шкале оценок при сравнительно малом числе заданий в тесте. Результат может быть получен за короткое время для большого числа участников. Интерпретация результатов проводится преимущественно на основе среднего арифметического балла тестируемых или рейтинга: для каждого испытуемого определяется процент тестируемых, у которых результат теста хуже или лучше его собственного [19]. При нормативно-ориентированном тестировании первоочередной задачей является определение не столько полноты освоения содержания обучения, сколько относительного места или рейтинга каждого из тестируемых, так как главная цель данного подхода – относительная дифференциация испытуемых по уровню подготовки. Сложность заданий такого теста лежит в широком диапазоне – от самых простых до очень сложных. Примерами нормативно-ориентированной интерпретации результатов педагогических измерений являются тесты ЕГЭ и централизованного (абитуриентского) тестирования.

Критериально-ориентированные тесты преследуют узкие конкретные цели, в большей степени соответствующие проверке, чем оценке: например, проверку уровня освоения определенного содержания обучения, выявление наличия определенного перечня знаний и умений, аттестацию выпускников на достижение ими минимально допустимого уровня компетентности по заданным критериям [20]. В отличие от нормативно-ориентированных критериально-ориентированные тесты обеспечивают дифференциацию только в сравнительно небольшой области вблизи порогово-

го балла. Для интерпретации результатов требуется относительно малое число заданий примерно одинаковой (критериальной) трудности, чтобы выявить, что знает и чего не знает испытуемый из тестируемого содержания обучения.

Содержательно-ориентированные тесты является инструментом выяснения степени усвоения каждым испытуемым отдельных элементов содержания учебной дисциплины. Для реализации содержательно-ориентированного подхода к организации и интерпретации результатов тестирования учебных достижений требуется большое число заданий, чтобы выводы о знании отдельных ДЕСО можно было точно детализировать.

Разные авторы [11, 14, 21–23], исходя из целеполагания разработки тестов, дополнительно классифицируют их следующим образом:

1) *диагностические* тесты, подразделяющиеся:

на тесты общих и специальных способностей;

на тесты обученности и учебных достижений;

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

3) тесты *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.

Важной характеристикой теста, обеспечивающей достоверность оценивания учебных достижений, корректность анализа результатов и их сопоставления, дифференцируемость оценок, является время решения теста. Оптимальное время тестирования определяется эмпирически и (или) экспертными оценками и указывается для каждого теста. Сложность теста определяется суммарной сложностью всех его заданий. Уровень сложности отдельных тестовых заданий и представительство в тесте заданий разной сложности определяется целями тестирования. Как правило, тесты формируются из заданий различной сложности, что обеспечивает дифференцирующую способность теста.

Традиционно мерой трудности отдельного j-го задания в тестологии считается доля правильных ответов p_j на него. В некоторых моделях тестов наряду с p_j в качестве показателя трудности задания применяется величина $q_j = 1 - p_j$ – доля неправильных ответов на j-е задание.

Мерой *уровня трудности заданий* в современной тестологии является *логит трудности задания*, определяемый натуральным логарифмом отношения доли неправильных ответов на задание, данных испытуемыми, к доле правильных ответов на данное 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 [24]. Сопоставление логарифмических оценок уровня знаний каждого испытуемого с уровнем трудности каждого задания посредством их вычитания позволяет создавать программно-инструментальные средства индивидуализации обучения и контроля, в том числе адаптивного.

Любые тесты обладают *тестологическими характеристиками*, т. е. признаваемыми профессиональным сообществом измерительными качествами. Считается, что тестологические характеристики могут быть оценены только апостериорно, например по итогам апробации на экспериментальной группе тестируемых. Такой постэмпирический подход к оцениванию характеристик теста практически отрицает прогнозирование их параметров априорно, на основе прогностических математических моделей и исходных данных. Эмпирические методы оценивания тестологических характеристик делают создание теста с заданными параметрами качества оценивания учебных достижений сложным, длительным и трудоемким процессом, связанным с выбором математической модели конструирования теста, определяющей его структуру, наполнением тестовой структуры заданиями соответствующего содержания и сложности, проведением апробационного тестирования, измерением искомых параметров характеристик теста и их сопоставлением с заданными, коррекцией теста и его дальнейшим совершенствованием методом последовательных

итераций до достижения соответствия заданным тестологическим свойствам.

Очевидным уязвимым местом рассмотренной итерационной процедуры конструирования теста и контроля параметров его характеристик является определение выборки тестируемых в процессе апробации. Во-первых, для получения состоятельных оценок выборка должна быть репрезентативной как по численности, так и по уровням учебных достижений или иных измеряемых параметров будущего контингента тестируемых. Для этого нужно как минимум иметь достоверную модель контингента, основанную на эмпирических данных (иное не рассматривается). Таким образом, налицо логическое противоречие: если есть достоверные априорные данные об измеряемых характеристиках (например, уровне учебных достижений) тестируемых, то ставится под сомнение необходимость разработки теста для их измерений. Если таких достоверных данных нет, то возникают обоснованные сомнения в корректности априорного соответствия выборки тестируемых для апробации основному контингенту тестируемых и, соответственно, достоверности результатов апробации. Если выборка делается для апробации теста из самого контингента тестируемых, то встает вопрос о корректности повторного использования той же выборки тестируемых для повторной апробации теста в ходе итерационной процедуры его конструирования. Если для повторной апробации выбирается другая часть тестируемого контингента, то с учетом требований к объему выборки, достаточному для ее репрезентативности, в апробации теста в ходе его итерационного конструирования может принять участие весь контингент тестируемых или его большая часть.

Другой уязвимой позицией данного подхода является неопределенность количественного описания характеристик теста, необходимого для задания конечных параметров, требующихся для конструирования теста, а также способов их измерения по итогам апробаций и критериев сравнения задаваемых и получаемых эмпирическим путем параметров, а также критериев сходимости (завершения) итерационных процедур конструирования теста.

Критический анализ постэмпирического подхода к разработке ТОУД и оцениванию его качества позволяет сделать вывод

о его нереализуемости на практике, по крайней мере в масштабах одной образовательной организации, в виде рассмотренной итерационной процедуры. В практике разработки тестов большую роль играют априорные экспертные оценки показателей качества теста, а также оценки на основе аналогий и опыта применения тестов самими разработчиками. Важную роль в обеспечении достоверности таких оценок призваны играть и математические модели оценивания характеристик теста, например надежности теста и его отдельных структурных компонентов (тестовых заданий, процедур формирования баллов и отображения на шкалы оценок), а также имитационные модели тестируемых и воспроизведения тестовых алгоритмов.

С другой стороны, апробирование теста и эмпирическое оценивание его характеристик на предмет его возможной корректировки, несомненно, способствует повышению достоверности последующих результатов тестирования. По нашему мнению, при разработке теста должны оптимально сочетаться априорные и апостериорные методы оценивания его качества. Первые, реализованные в математических моделях и экспертных оценках, существенно снижают трудоемкость и время разработки теста. Вторые позволяют окончательно и предметно судить о достижении поставленных разработчиками целей.

Важной характеристикой теста является *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 свидетельствуют либо об одинаковом уровне знаний и умений тестируемых, либо о несостоятельности теста в дифференцировке знаний и умений испытуемых. Удобной 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.

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.

Требуемая точность педагогических измерений заданиями теста теоретически достигается именно надежностью теста. Это вытекает из известного постулата о неизбежности погрешности любых измерений: измеряемая величина X не равна истинному значению T [11]. 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 [25].

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 главе 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. Анастази, С. Урбина [26], Ю.М. Неймана, В.А. Хлебникова [24], М.Б. Челышковой [27]. 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 вычисляется из соотношения: $r_t' = \dfrac{2r_t}{1+r_t}$.

Валидность – характеристика способности теста оценивать требуемые показатели. Очень часто процесс создания теста носит многоцелевой характер, поэтому валидность оценивают с разных

позиций и руководствуются различными критериями целевой адекватности теста.

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 [28]. Валидность оценивается на основе разнообразной информации и подразделяется на несколько типов:

– диагностическую, отражающую способность теста дифференцировать испытуемых по изучаемому признаку, например по результатам тестирования структурировать знания и умения испытуемых;

– 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» [25];

– эмпирическую, определяемую экспертными оценками специалистов (педагогов, ученых-экспертов, сотрудников центра тестирования и др.);

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

– содержательную, определяемую адекватностью заданий проверяемому содержанию обучения.

К указанным выше типам валидности В.С. Аванесов, В.П. Беспалько, И.П. Подласый добавляют функциональную и критериальную валидности.

Функциональная валидность определяет соответствие задания уровню усвоения контролируемых знаний. Критериальная валидность отражает направленность теста на измерение соответствия учебных достижений заранее определенным критериям или эталонам, например образовательному стандарту. Количественной мерой критериальной валидности служат коэффициенты корреляции между показателями теста и заданной критериальной мерой.

1.2. Разновидности тестовых заданий

Формы тестовых заданий подразделяются на две группы: задания в открытой форме (открытого типа) и задания в закрытой форме (закрытого типа).

В тестовых заданиях открытого типа набор готовых ответов для выбора отсутствует, испытуемым необходимо запрашиваемую информацию вводить самостоятельно. Задания открытого типа подразделяются на задания с дополнением и задания свободного изложения. В обоих случаях испытуемым необходимо самостоятельно дополнить содержание задания. В первом случае дополнение является кратким: как правило, необходимо ввести от одного до трех слов или символов. В результате задание должно превратиться в истинное логическое высказывание или формулу. 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.

На практике чаще используются задания закрытого типа. В таких заданиях дополнительная информация испытуемыми не вводится, необходимо выбрать один или несколько ответов из предложенных вариантов или каким-либо образом упорядочить лингвистические или символьные конструкции. Варианты неверных ответов в тестовых заданиях с выбором одного или нескольких правильных ответов называются *дистракторами*. Количество дистракторов должно обеспечивать низкие значения вероятности случайного угадывания правильного ответа. С другой стороны, большое количество дистракторов приводит к увеличению времени решения задания и повышает вероятность случайной ошибки при выборе правильного ответа. Как правило, в заданиях с выбором единственно верного ответа количество дистракторов составляет три или четыре. Их меньшее количество обычно обусловлено спецификой задания, как, например, в тестах на знание правил дорожного движения. Дистракторы должны отвечать принципам правдоподобия и равной привлекательности. Авторы настоящей монографии дополнили эти требования принципом стилистической однородности (раздел 4.3).

Ряд специалистов полагают, что выбор дистракторов должен иметь равномерное распределение, т. е. варианты неверных ответов должны выбираться с одинаковой частотой. По нашему мнению, анализ неправильных ответов испытуемого представляет для педагога не меньший интерес, чем анализ соотношения верных и неверных решений тестовых заданий. Так, 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. 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. В любом случае 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, по нашему мнению, поможет не только избежать завышения оценок тестирования учебных достижений, но и повысить дифференцирующую способность теста.

Очевидно, что вероятность выбора дистракторов будет уменьшаться с увеличением 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 его отдельных заданий и необходимости модернизации базы заданий. С целью повышения дифференцирующей способности задания дистрактор, который никто не выбирает в качестве правильного ответа, необходимо откорректировать или удалить. Считается, что творчество разработчика теста состоит именно в создании неправильных, но очень правдоподобных ответов.

М.Б. Челышкова рекомендует для создания дистракторов предъявлять обучающимся задания в открытой форме. Последующий анализ ошибок в представленных обучающимися ответах дает возможность создать правдоподобные, с точки зрения испытуемых, дистракторы [20].

В.С Аванесов в [9] выделяет две группы принципов формулирования заданий: для подбора ответов к заданиям и для разработки содержания заданий.

Принципы подбора ответов к заданиям с выбором ответа:

1. Противоречивости – ответы подбираются взаимоисключающими, с использованием отрицания «не».

Пример. При изотермическом расширении газа его давление:

а) уменьшается;

б) не уменьшается.

2. Противоположности – ответы формулируются как антонимы, допускаются переходные варианты от правильного ответа к взаимоисключающему.

Пример. При изотермическом расширении газа его давление:

а) понижается;

б) повышается.

3. Однородности – варианты ответа относятся к одному множеству, смысловому или гомологическому ряду.

Пример. Изобара на термодинамической диаграмме отображает процесс, происходящий при постоянстве:

а) давления;

б) объема;

в) температуры.

4. Кумуляции – включения предыдущих ответов в последующие с помощью союза «и» и запятых.

Пример. Изобара на термодинамической диаграмме отображает процесс, происходящий при постоянстве:

а) давления;

б) давления и температуры;

в) давления, температуры и объема.

5. Сочетания – использованя в вариантах ответа сочетаний одинаковых или близких по смыслу слов, знаков, терминов и др.

Пример. При изобарном процессе изменяются:

а) объем и температура;

б) объем и давление;

в) температура и давление.

6. Градуирования – использования градаций какой-либо характеристики для заданий с тремя и большим числом вариантов ответа.

Пример. При изотермическом расширении газа его давление:

а) понижается;

б) не изменяется;

в) повышается.

7. Удвоенного противопоставления – использования отрицания двух параметров для заданий с четырьмя ответами.

Пример. При изотермическом расширении газа:

а) уменьшается давление и плотность;

б) уменьшается давление и не меняется плотность;

в) уменьшается плотность и не меняется давление;

г) не меняется ни давление, ни плотность.

Содержательная основа заданий 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. Такой подход снижает вероятность искажения результатов, возникающих за счет списывания, и тем самым повышает объективность сопоставления результатов испытуемых.

Пример. При изотермическом (изобарическом) расширении газа уменьшается (увеличивается, не изменяется) его давление (температура).

В данном утверждении сформировано три фасета, из которых могут создаваться различные варианты тестового задания.

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

Пример. Если газ расширяется при постоянной температуре, то уменьшается его:

а) давление;

б) молярная масса;

в) молярный объем.

Задания закрытого типа подразделяются на следующие виды:

– с альтернативными ответами;

– с выбором единственного верного ответа;

– с множественным выбором верных ответов;

– на установление соответствий;

– на установление последовательности;

– с градуированными ответами.

Задания с альтернативными ответами являются самыми простыми тестовыми заданиями закрытого типа. Они, как правило, формулируются в виде согласия или несогласия с утверждением.

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

а) обязан остановиться;

б) не обязан останавливаться.

Задания такого рода в тестах используются редко, поскольку вероятность случайного угадывания правильного ответа испытуемыми является очень высокой, равной 0,5. Их рекомендуется использовать для быстрого и грубого отсева испытуемых по принципу «зачет/незачет». Иногда с целью уменьшения вероятности случайного угадывания, для проверки знания одной ДЕСО применяют серии заданий с альтернативами. При таком подходе 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.

В *задании с выбором единственно верного ответа* испытуемому предлагается несколько вариантов ответа, среди которых только один верный. Структура данных заданий 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 заданий с выбором единственно верного ответа являются:

– ограниченность диапазона тестируемых знаний и умений (как правило, тестовые задания данного типа не применяются для оценивания комплексных знаний и умений, умений творческого характера и т. п.);

– относительно высокая вероятность угадывания испытуемым правильного ответа.

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

В *заданиях с множественным выбором верных ответов* испытуемый должен в представленном перечне вариантов ответа указать все верные.

Пример. Полностью на территории России расположена река:

а) Волга;

б) Иртыш;

в) Урал;

г) Енисей;

д) Амур;

е) Ангара.

Задания данного типа характеризуются существенно меньшей вероятностью угадывания полностью верного ответа. Однако оценивание выполнения таких заданий является менее однозначным по сравнению с заданиями с единственным выбором.

В.С. Аванесов предлагает за полностью правильное решение давать три балла, за каждую ошибку отнимать один балл [9]. В случае допущения испытуемым более трех ошибок значение имеющихся баллов не изменяется. Таким образом, в результате выполнения единичного задания количество возможных баллов варьируется от 0 до 3. Предложенная схема позволяет получать только положительные баллы в предположении, 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 [20]. 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 [29].

Разные способы начисления баллов влекут за собой различные методики перевода баллов в привычные оценки. В выборе диапазонов оценок и критериев соответствия неизбежен субъективизм разработчиков тестов, применение эвристических и эмпирических подходов, что придает этапу апробации тестов на основе заданий с множественным выбором верных ответов особое значение и требует принятия дополнительных мер по обеспечению репрезентативности экспериментальной выборки тестируемых.

Задания с множественным выбором верных ответов менее универсальны по сравнению с заданиями с выбором единственно верного ответа и требуют более высокой квалификации разработчиков тестов.

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

Структурно задание на установление соответствия 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. Укажите столицы государств Южной Америки:

Эквадор	Монтевидео
Перу	Кито
Боливия	Каракас
Венесуэла	Лима
Уругвай	Сукре

Существует альтернативная форма заданий на установление соответствия – задания с множественным соответствием. В таких заданиях каждому элементу левого столбца могут соответствовать несколько элементов правого столбца, отдельным элементам левого столбца может не находиться соответствий.

Еще одна форма заданий на установление соответствия предполагает наличие третьего столбца. В таких заданиях элементы первого множества сопоставляются с элементами второго и третьего множеств.

В обоих случаях, как и в заданиях с множественным выбором верных ответов, возникает неоднозначность трактовок верных решений и процедур начисления баллов за исчерпывающие и частично верные ответы.

Частным случаем заданий на установление соответствия с точки зрения логики их конструирования являются *задания на установление правильной последовательности*. В таких заданиях испытуемому необходимо расположить в нужной последовательности элементы ответа. С помощью заданий на установление правильной последовательности удобно проверять знание и понимание испытуемыми формулировок определений, понятий, терминов.

Пример. 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 цифр или, если позволяет интерфейс, путем перестановки элементов ответа до образования последовательно 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 заданий с множественным выбором верных ответов, возникает проблема неоднозначности оценивания.

Задания с градуированными ответами или выбором наилучшего ответа содержат ответы, которые имеют градацию по степени правильности. Задача составителя заключается в том, чтобы найти и применить признак, позволяющий осуществить такую градацию. Максимальное количество баллов начисляется, если градация ответов полностью совпадает с эталонной после-

довательностью, определяемой разработчиком задания. Случайное угадывание верных решений таких заданий маловероятно.

Пример. Относительная ширина полосы частот резонансного контура определяется значениями:

1) добротности контура;

2) емкости и индуктивности;

3) индуктивности и сопротивления;

4) емкости.

Конструирование такого рода заданий 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 заданий. Для получения правильного решения испытуемый должен не просто знать тестируемый 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 задания значительно упрощает составление задания и делает более прозрачными критерии его верного решения. Такой перечень можно формировать с применением принципов кумуляции и сочетания, испытуемый должен градуировать ответы по степени полноты и достаточности перечня признаков и свойств объектов. Однако в этом случае задание с градуированными ответами практически вырождается в задание с множественным выбором правильных ответов.

Пример. 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.

В главе 4 нами предложена структура тестового задания, сочетающего в себе достоинства заданий с градуированными ответами и выбором единственного правильного ответа, существенно расширяющая область задач с возможностью градации ответов.

1.3. Статистический анализ обработки результатов тестирования

Статистический анализ обработки результатов тестирования позволяет объективно определить результаты испытуемых, оценить качество тестовых заданий и надежность теста. Большой вклад в статистическое обоснование качества теста внесли Г. Гулликсен, В.С. Аванесов, А.Н. Майоров, М.Б. Челышкова, В.С. Ким и др.

Для выполнения статистической обработки результатов тестирования первоначально формируют *бинарную матрицу* тестовых результатов, элементы которой принадлежат множеству {0,1}: нулю соответствует неверный ответ, единице – верный ответ.

В качестве примера рассмотрим бинарную матрицу, полученную в результате тестирования [30].

Таблица 1.1

Бинарная матрица

Испытуемые		Номера тестовых заданий								
№	ФИО	1	2	3	4	5	6	7	8	9
1	Алексеев Л.А.	0	1	1	0	0	1	0	1	1
2	Антонов К.Е.	1	0	1	0	1	1	0	0	0
3	Арбузов В.С.	1	0	1	1	0	0	1	1	0
4	Богатырев И.Ф.	0	1	1	1	0	1	0	1	0
5	Васнецов П.С.	0	0	1	0	1	0	0	0	0
6	Иванов А.В.	1	1	1	1	1	1	1	1	1
7	Кузнецов Ю.Д.	0	0	1	0	1	1	1	0	0
8	Петров С.Ю.	1	1	1	1	0	0	1	1	0
9	Серов В.И.	0	0	1	0	1	1	0	0	1
10	Сидоров Д.К.	1	0	1	1	1	1	0	0	1
11	Яковлев О.В.	0	1	1	0	0	0	1	1	0

Пусть X_i – значения индивидуального балла *i*-го испытуемого; R_j – количество верных ответов на *j*-е задание; W_j – количество неверных ответов на *j*-е задание; p_j – доля верных ответов; q_j – доля неверных ответов.

$$X_i = \sum_{j=1}^{M} a_{ij}\ ,\ i = 1,2,...,N\ ,$$

где M – количество тестовых заданий.

$$R_j = \sum_{i=1}^{N} a_{ij}\ , j = 1,2,...,M\ ,$$

где N – число испытуемых.

$$W_j = N - R_j\ , j = 1,2,...,M\ .$$

Рассчитаем индивидуальный тестовый балл для второго испытуемого:

$$X_2 = \sum_{j=1}^{M} a_{2j} = 4.$$

Вычислим количество верных и неверных ответов для четвертого задания:

$$R_4 = \sum_{i=1}^{N} a_{i4} = 5, \quad W_4 = N - R_4 = 11 - 5 = 6\ .$$

Доли верных ответов (мера трудности тестового задания) p_i и неверных ответов q_j на *j*-е задание:

$$p_j = \frac{R_j}{N}\ ,\ j = 1,2,...,M\ ;\ q_j = 1 - p_j\ .$$

Для четвертого задания получим:

$$p_4 = \frac{5}{11} \approx 0{,}4545; \quad q_4 = 1 - 0{,}4545 = 0{,}5455\ .$$

Для каждого испытуемого и для каждого тестового задания вычислим значения X_i и R_j. Для удобства анализа фамилии испытуемых заменим их номерами и произведем ранжирование по X_i и R_j.

Таблица 1.2

Бинарная матрица, ранжированная по величинам X_i и R_j

№	3	6	5	8	1	2	4	7	9	X_i
6	1	1	1	1	1	1	1	1	1	9
8	1	0	0	1	1	1	1	1	0	6
10	1	1	1	0	1	0	1	0	1	6
1	1	1	0	1	0	1	0	0	1	5
3	1	0	0	1	1	0	1	1	0	5
4	1	1	0	1	0	1	1	0	0	5
2	1	1	1	0	1	0	0	0	0	4
7	1	1	1	0	0	0	0	1	0	4
9	1	1	1	0	0	0	0	0	1	4
11	1	0	0	1	0	1	0	1	0	4
5	1	0	1	0	0	0	0	0	0	2
R_j	11	7	6	6	5	5	5	5	4	54

Анализ полученной матрицы показывает, что с третьим заданием справились все испытуемые. Данное задание не позволяет дифференцировать испытуемых, его следует удалить из теста. Первую строку в таблице также следует удалить, так как испытуемый № 6 правильно выполнил все задания теста.

Таблица 1.3

Редуцированная бинарная матрица

№	6	5	8	1	2	4	7	9	X_i
8	0	0	1	1	1	1	1	0	5
10	1	1	0	1	0	1	0	1	5
1	1	0	1	0	1	0	0	1	4
3	0	0	1	1	0	1	1	0	4
4	1	0	1	0	1	1	0	0	4
2	1	1	0	1	0	0	0	0	3
7	1	1	0	0	0	0	1	0	3
9	1	1	0	0	0	0	0	1	3
11	0	0	1	0	1	0	1	0	3
5	0	1	0	0	0	0	0	0	1
R_j	6	5	5	4	4	4	4	3	35

Важным параметром тестового задания является вариация (дисперсия) тестовых баллов $p_j q_j$. Чем выше значение вариации,

тем лучше задание дифференцирует испытуемых. Дополним матрицу значениями p_j, q_j, $p_j q_j$.

Таблица 1.4

Редуцированная бинарная матрица
с дополнительными параметрами

№	6	5	8	1	2	4	7	9	X_i	X^2_i
8	0	0	1	1	1	1	1	0	5	25
10	1	1	0	1	0	1	0	1	5	25
1	1	0	1	0	1	0	0	1	4	16
3	0	0	1	1	0	1	1	0	4	16
4	1	0	1	0	1	1	0	0	4	16
2	1	1	0	1	0	0	0	0	3	9
7	1	1	0	0	0	0	1	0	3	9
9	1	1	0	0	0	0	0	1	3	9
11	0	0	1	0	1	0	1	0	3	9
5	0	1	0	0	0	0	0	0	1	1
R_j	6	5	5	4	4	4	4	3	35	135
W_j	4	5	5	6	6	6	6	7		
p_j	0,6	0,5	0,5	0,4	0,4	0,4	0,4	0,3		
q_j	0,4	0,5	0,5	0,6	0,6	0,6	0,6	0,7		
$p_j q_i$	0,24	0,25	0,25	0,24	0,24	0,24	0,24	0,21		

Построим функциональную зависимость вариации тестовых баллов от трудности задания (рис. 1.1).

Рис. 1.1. Вариация задания

Максимальное значение $p_j q_j = 0{,}25$ достигается при $p_j = 0{,}5$. При $p_j = 0$ (верный ответ на задание не получен ни от одного испытуемого) и $p_j = 1$ (все испытуемые успешно ответили на вопрос) дисперсия задания равна нулю.

Таблица 1.4 имеет характерную особенность – большая часть нулей и единиц распределились относительно диагонали, идущей из левого нижнего угла в правый верхний. По Л. Гуттману [31], это разграничение должно быть идеальным: если испытуемый верно ответил на трудное задание, то он должен справиться и с более легкими заданиями. В действительности же это правило работает далеко не всегда. Например, в нашем примере восьмой испытуемый верно выполнил трудное задание № 7, но не справился с более простыми заданиями № 5 и № 6. Профиль испытуемого искажен.

Отклонения от идеального разграничения могут свидетельствовать о неверной структуре знаний испытуемого, нарушениях процедуры тестирования или недостатках тестовых заданий.

Данные, представленные в таблице 1.4, удобно анализировать, используя их графическое представление. Составим вариационный ряд, где X_i – значения индивидуальных баллов испытуемых, n_i – частоты индивидуальных баллов.

Таблица 1.5

Дискретный вариационный ряд

X_i	1	3	4	5
n_i	1	4	3	2

Построим полигон частот (рис. 1.2.).

Рис. 1.2. Полигон частот

Вычислим основные характеристики дискретного статистического ряда: моду, медиану, среднее значение.

Мода – значение, имеющее наибольшую частоту: $M_0 = 3$.

Медиана – значение, приходящееся на середину ранжированной выборки:

$$M_e = \frac{3+4}{2} = 3{,}5 .$$

Среднее значение:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i n_i = \frac{1\cdot 1 + 3\cdot 4 + 4\cdot 3 + 5\cdot 2}{10} = 3{,}5 .$$

Нормативно-ориентированный тест должен хорошо дифференцировать испытуемых. Это означает, что индивидуальные тестовые баллы должны в достаточной степени отличаться друг от друга.

Вариацию тестовых результатов определяют отклонения от среднего значения: $\Delta = \left(x_i - \bar{x}\right)$.

При полном совпадении всех индивидуальных баллов вариация равна нулю. Если индивидуальные баллы не совпадают, то отклонения могут быть положительными и отрицательными. Сумма всех отклонений будет равна нулю. Поэтому, чтобы охарактеризовать вариацию тестовых баллов, используют квадрат отклонений. Сумма квадратов отклонений зависит от количества испытуемых. Таким образом, для дальнейшего анализа результатов тестирования необходимо вычислить следующие числовые характеристики: дисперсию (D), среднее квадратическое отклонение (σ), исправленную дисперсию (S^2), исправленное среднее квадратическое отклонение (S).

$$D_в = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^2 \cdot n_i = \frac{(1-3{,}5)^2\cdot 1 + (3-3{,}5)^2\cdot 4 + (4-3{,}5)^2\cdot 3 + (5-3{,}5)^2\cdot 2}{10} = 1{,}25 .$$

$$\sigma_в = \sqrt{D_в} = \sqrt{1{,}25} \approx 1{,}118 .$$

$$S^2 = \frac{n}{n-1} D_в = \frac{10}{10-1}\cdot 1{,}25 \approx 1{,}3889 .$$

$$S = \sqrt{S^2} = \sqrt{1{,}3889} \approx 1{,}1785 .$$

Величина дисперсии тестовых баллов позволяет судить о дифференцирующей способности теста и качестве тестирования в целом. Малая величина дисперсии говорит о том, что тест плохо различает испытуемых по уровню знаний, не позволяет с прием-

лемой точностью ранжировать их. Слишком большая дисперсия указывает на сильную неоднородность группы испытуемых, на возможные нарушения процедуры тестирования, на недостаточно ясные формулировки заданий и т. п. В случае оптимального значения дисперсии распределение тестовых баллов близко к нормальному.

Непрерывная случайная величина X имеет нормальный закон распределения (закон Гаусса) с параметрами a и σ^2, если ее плотность вероятности имеет вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad (1.1)$$

где a – математическое ожидание (среднее арифметическое), σ^2 – дисперсия.

Рис. 1.3. Распределение Гаусса (нормальное распределение) с параметрами $a=0,\ \sigma^2=1{,}5$, $a=0,\ \sigma^2=1$, $a=0,\ \sigma^2=0{,}8$, $a=1,\ \sigma^2=1$.

Для сравнения результатов тестирования с (1.1) используются дополнительные характеристики вариационного ряда: асимметрия и эксцесс.

Начальным выборочным моментом порядка k называется среднее арифметическое k-х степеней всех значений выборки:

$$\widetilde{\nu}_k=\frac{\sum_{i=1}^{m}x_i^k\cdot n_i}{n}. \qquad (1.2)$$

Центральным выборочным моментом порядка k называется среднее арифметическое k-х степеней отклонений наблюдаемых значений выборки от выборочного среднего $\overline{x_в}$:

$$\widetilde{\mu}_k = \frac{\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^k \cdot n_i}{n}. \qquad (1.3)$$

Выборочный коэффициент асимметрии вычисляется по формуле:

$$\widetilde{A} = \frac{\widetilde{\mu}_3}{\sigma_в^3}. \qquad (1.4)$$

Характеризует асимметрию полигона вариационного ряда. Если $\widetilde{A} = 0$, то распределение имеет симметричную форму, т. е. варианты, равноудаленные от x, имеют одинаковую частоту (среднее арифметическое, мода и медиана совпадают). Если это равенство нарушается, то распределение асимметрично.

Если $\widetilde{A} < 0$, то наблюдается положительная (правосторонняя) асимметрия, т. е. более пологий «спуск» полигона слева. Если $\widetilde{A} > 0$, то наблюдается отрицательная (левосторонняя) асимметрия, т. е. более пологий «спуск» полигона справа.

Выборочный коэффициент эксцесса или *коэффициент крутости* вычисляется по формуле:

$$\widetilde{E} = \frac{\widetilde{\mu}_4}{\sigma_в^4} - 3. \qquad (1.5)$$

Характеризует крутизну вершин графика выборочного распределения по сравнению с графиком нормального распределением. Коэффициент эксцесса для случайной величины, распределенной по нормальному закону, равен нулю.

Если $\widetilde{E} > 0$ ($\widetilde{E} < 0$), то полигон вариационного ряда имеет более крутую (пологую) вершину по сравнению с нормальной кривой.

Для нашего примера центральный выборочный момент 3-го порядка:

$$\widetilde{\mu}_3 = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^3 \cdot n_i = \frac{(1-3{,}5)^3 \cdot 1 + (3-3{,}5)^3 \cdot 4 + (4-3{,}5)^3 \cdot 3 + (5-3{,}5)^3 \cdot 2}{10} = -0{,}9.$$

Центральный выборочный момент 4-го порядка:

$$\widetilde{\mu}_4 = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \overline{x_в}\right)^4 \cdot n_i = \frac{(1-3{,}5)^4 \cdot 1 + (3-3{,}5)^4 \cdot 4 + (4-3{,}5)^4 \cdot 3 + (5-3{,}5)^4 \cdot 2}{10} = 4{,}9625.$$

Коэффициент асимметрии:

$$\widetilde{A} = \frac{\widetilde{\mu}_3}{\sigma_в^3} = \frac{-0{,}9}{1{,}118^3} \approx -0{,}644 .$$

Коэффициент эксцесса:

$$\widetilde{E} = \frac{\widetilde{\mu}_4}{\sigma_в^4} - 3 = \frac{4{,}9625}{1{,}118^4} - 3 = 0{,}176 .$$

Так как коэффициент асимметрии $\widetilde{A}$ отрицателен, то наблюдается левосторонняя асимметрия. Поскольку коэффициент эксцесса $\widetilde{E}$ положителен, то рассматриваемое распределение имеет более крутую вершину по сравнению с нормальной кривой.

График нормального распределения (кривая Гаусса) и эмпирическая кривая (полигон относительных частот статистического распределения) приведены на рис. 1.4.

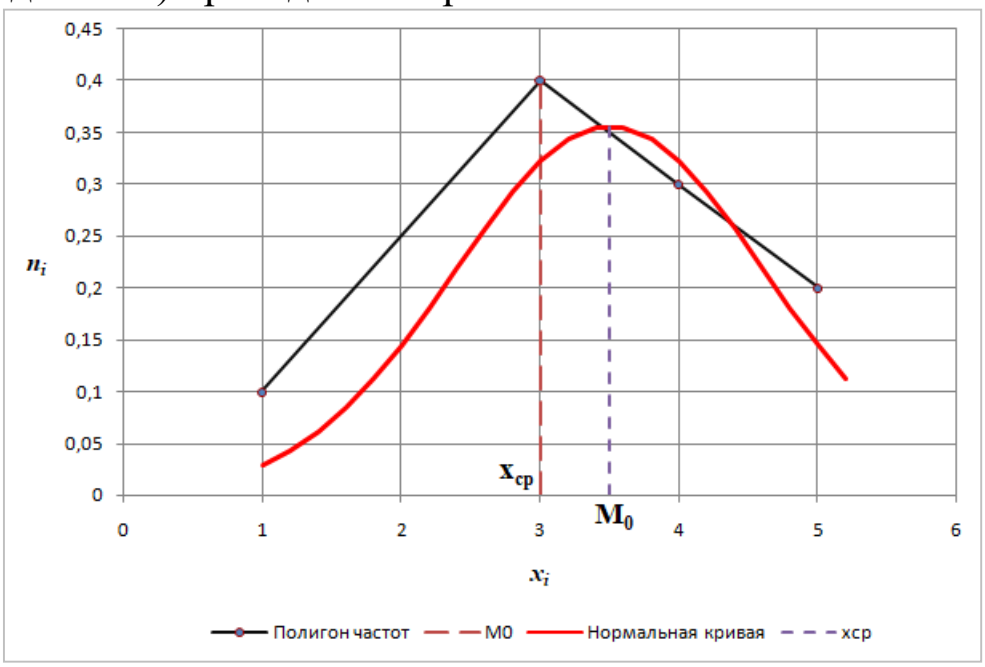


Рис. 1.4. Нормальное распределение и эмпирическая кривая

2. Модель оценки качества компьютерных тестов для проверки знаний и умений обучающихся

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

Благодаря своим дидактическим возможностям и доступности программных оболочек с удобным интерфейсом компьютерные тесты занимают прочные позиции среди инструментов контроля знаний и умений обучающихся и давно уже вышли из категории нетрадиционных для отечественной педагогики дидактических средств. В то же время, несмотря на безусловный прогресс в совершенствовании методик разработки тестов и возрастание объемов компьютерного тестирования, обусловленное расширенным внедрением дистанционных образовательных технологий в реализацию программ высшего и дополнительного профессионального образования, компьютерные тесты, как правило, играют вспомогательную роль в системе контроля знаний, выполняя значительный объем «черновой» работы при формировании оценок предварительной и промежуточной аттестации, контроля текущей успеваемости и остаточных знаний и оставляя последнее слово за методиками индивидуального опроса, беседы, письменных контрольных работ. Подтверждением этому отчасти является и сокращение объема и значимости тестов с выбором единственного верного ответа в материалах и алгоритмах формирования оценок ЕГЭ.

Очевидно, степень доверия педагогов к тестам напрямую связана с достоверностью результатов тестирования, т. е. адекватностью соотнесения оценок, полученных по итогам тестов, с реальными знаниями и умениями испытуемых. В идеале следует рассматривать тождественность оценок тестов $О(T_т)$ в момент времени $T_т$ неким безошибочным, абсолютно достоверным оценкам знаний и умений обучающихся $О^*(T_о)$, полученным независимыми способами в период времени оценивания:

$$T_о \in T_т \pm \Delta T_и, \qquad (2.1)$$

где $\Delta T_{\text{и}}$ – временной интервал, в течение которого возможно подлежащее учету (например, 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 [13, 32].

Если предположить, что оценки $О^*(T_{\text{о}})$ характеризуются некоторой абсолютной погрешностью:

$$\Delta О^*(T_{\text{о}}) = О^*(T_{\text{о}}) - О_{\text{и}}(T_{\text{о}}), \quad (2.2)$$

где $О_{\text{и}}(T_{\text{о}})$ – некая идеальная оценка учебных достижений обучающихся, интерпретируемая, например, как математическое ожидание частных оценок $О_i^*(T_{\text{о}})$, полученных наиболее достоверным способом оценивания:

$$О_{\text{и}}(T_{\text{о}}) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} О_i^*(T_{\text{о}}), \quad (2.3)$$

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

$$О_{\text{т}}(T_{\text{о}}) \in [О_i^*(T_{\text{о}}) - \Delta О^*(T_{\text{о}}); О_i^*(T_{\text{о}}) + \Delta О^*(T_{\text{о}})]. \quad (2.4)$$

В качестве дополнительного условия может рассматриваться неравенство:

$$\Delta О^*(T_{\text{о}}) < \Delta О_{\text{т}}(T_{\text{о}}), \quad (2.5)$$

где

$$\Delta О_{\text{т}}(T_{\text{о}}) = О_{\text{т}}(T_{\text{о}}) - О_{\text{и}}(T_{\text{о}}). \quad (2.6)$$

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

$$О_{\text{т}}(T_{\text{о}}) \pm \Delta О_{\text{т}}(T_{\text{о}}) \in [О_i^*(T_{\text{о}}) - \Delta О^*(T_{\text{о}}); О_i^*(T_{\text{о}}) + \Delta О^*(T_{\text{о}})], \quad (2.7)$$

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, 33, 34, 35]. Однако для разработчика теста важно оценить именно потенциальную достоверность оценок учебных достижений обучающихся априори, на основе прогностических методик, базирующихся на математических моделях, как описывающих механизмы формирования оценок, так и дающих количественные оценки показателей, способных характеризовать априорную достоверность результатов тестирования.

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

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

– ограниченностью времени разработки теста $\tau_р$, его апробации $\tau_а$ и собственно самого контроля учебных достижений обучающихся $\tau_к$ значениями $\Delta T_и$:

$$\tau_р + \tau_а + \tau_к < \Delta T_и; \qquad (2.8)$$

– необходимостью априорного выбора оптимальных решений в отношении структуры теста (количества и вида тестовых заданий, дифференциации уровней их сложности), 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 заданий, или существенным увеличением общей трудоемкости разработки теста при исключении применявшихся на этапе его апробации заданий из их базы и их замены на эквивалентные. Такого рода замена может поставить под сомнение в целом корректность оценки качества теста, поскольку степень эквивалентности (по содержанию, сложности, дифференцирующей способности и др.) исключаемых из базы и вводимых вместо них заданий определяется разработчиком теста субъективно.

Для оценки (как априорной – потенциальной, прогнозируемой, так и апостериорной – эмпирической) достоверности теста в

педагогической литературе используются такие характеристики, как «эффективность», «пригодность», «валидность» и «надежность» [7, 32].

Руководствуясь стандартизированным определением эффективности как «соотношением между достигнутым результатом и использованными ресурсами» [36], следует рассматривать эффективность теста как отношение некого количественного показателя, характеризующего достигнутый результат, к количественному показателю, характеризующему трудоемкость или стоимость получения данного результата. Если с определением величины знаменателя не возникает принципиальных проблем и в его вычислении можно 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., то в вычислении числителя имеется значительная неопределенность. Прежде всего в публикациях, упоминающих эффективность применительно к тестам, предназначенным для оценки учебных достижений обучающихся, не оговаривается достигаемый или желаемый эффект, в качестве которого могут рассматриваться как собственно полученные при тестировании оценки знаний и умений, так и приемлемые значения отдельных характеристик достоверности полученных результатов или положительные приращения значений этих характеристик. Ввиду отсутствия соотнесения каких-либо достигаемых или улучшаемых количественных характеристик теста с затраченными на достижение результата ресурсами следует предположить, что 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 его достоверность оценивания.

Согласно В.С. Аванесову, эффективное тестирование – это обязательно индивидуализированное измерение уровня подго-

товки каждого испытуемого с помощью теста, оптимального по трудности и минимального по количеству заданий [37]. Из данного определения следует, что при некотором фиксированном результате, например при обеспечении приемлемой достоверности оценок теста, повышение эффективности теста может быть достигнуто за счет снижения количества заданий. Однако в этом случае остается открытым вопрос оптимальности трудности теста. Следует отметить и то, что оптимизация трудности теста и одновременная минимизация количества заданий – задача повышенной сложности для конструктора теста, и ее решение неизбежно увеличивает ресурсоемкость разработки теста, которая должна быть учтена в оценке эффективности.

Пригодность как свойство теста, по нашему мнению, является проекцией на частные методики разработок и применения тестов понятий «валидность» и «достоверность». Проблемой оценки данного свойства является необходимость обоснования одного или нескольких количественных показателей, его характеризующих, и формирования соответствующих оценочных шкал с разработкой методик их применения. Без решения данной проблемы пригодность 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: «пригоден» – «не пригоден», что делает перспективы ее применения для оценки качества тестов сомнительными.

Проблемой применения термина «валидность» в качестве свойства теста для контроля знаний 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. 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, 2]. Необходимо учесть, что вопрос валидности возникает при определении количественных или качественных величин по-

средством косвенных измерений, результаты которых связаны с определяемой величиной неявным или неоднозначным образом, например определение степени ожирения человека по его росту и массе тела. Зачастую валидность методик измерений относится к выявлению, количественному или качественному описанию и анализу латентных величин. Ее обоснование во всех случаях 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 теста, уровнем сложности заданий и алгоритмом подсчета баллов регулируется непосредственно разработчиком теста исходя из дидактических целей и частных задач выявления уровня учебных достижений обучающихся.

Следует также учесть, что 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. прошлого века, 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»). В итоге буквальное привнесение термина «валидность» из психологической квалиметрии, где верификация результатов тестирования, как правило, осложняется отсутствием эталонных измерений оцениваемых качеств и, вследствие невысокой достоверности оценок, получаемых независимыми способами, зачастую осуществляется повторными (в том числе многократно) 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 как характеристики достоверности оценок или свойства качества теста сопряженным с проблемами методологического характера.

Так, определения валидности оперируют либо понятиями высокой степени абстрактности, 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 («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 [38].

В. С. Аванесов утверждает, 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» [11]. Однако ни один из элементов данного перечня, ни их совокупность не могут быть рассмотрены в качестве меры, характеризующей степень валидности.

Для конкретизации данного, столь неоднозначно трактуемого различными исследователями, понятия в [39–42] введены более частные понятия: «содержательная (в том числе внешняя) ва-

лидность», «критериальная (статистическая, прогностическая) валидность» «конструктная (конструктивная) валидность», «эмпирическая валидность»), в определениях которых тем не менее отсутствуют указания не только на способы, но и на возможности количественного описания каких-либо характеризующих их показателей. Наиболее «прозрачным» в отношении количественного описания из перечисленных терминов является эмпирическая валидность – «независимый показатель, в котором используются экспертные оценки и характеристики теста, данные специалистами», подразумевающий хотя бы способ получения оценок – экспертный опрос независимых специалистов, опирающийся на определенные методики получения и обработки экспертных данных [7, 43], однако не определяющий совокупность оцениваемых показателей и критерии валидности теста в целом на основании полученных оценок.

Нередко вновь вводимые термины содержат в себе другие неопределенные или неоднозначные понятия, например: «конструктная 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» [44]. Не проясняет ситуацию с количественными измерениями и их анализом на соответствие заявляемому уровню качества теста и такое пояснение: «конструктивная валидность используется при сложности или невозможности подобрать адекватные критерии валидизации. При этом используется комплекс характеристик, свидетельствующих о теоретической обоснованности методики, соответствии полученных с помощью теста результатов теоретическим ожиданиям и закономерностям [1]».

Еще более осложняют применение валидности как комплексной характеристики качества теста для оценки учебных достижений обучающихся следующие ее типы и подвиды:

– диагностическая (конкурентная) и прогностическая [25];

– очевидная и концептуальная [25, 42];

– функциональная и критериальная [25, 35].

Это позволяет согласиться с утверждениями исследователей о том, что «в понятие валидности входит самая разнообразная информация о тесте».

Наиболее близкое приближение к представлению валидности количественными характеристиками сделано В. Пугачевым, который утверждает, что «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» [45].

Однако коэффициент корреляции применительно к тесту означает корреляцию данных, полученных в разных условиях (например, для разных выборок тестируемых) по одному и тому же тесту. В приведенном выше способе количественной оценки валидности корректнее говорить о коэффициенте корреляции результатов (статистических данных), полученных несколькими способами, включая рассматриваемый тест. Здесь же используется новый термин «добротность», очевидно, синонимичный достоверности и валидности. Тем не менее является очевидным 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 [46].

Применительно к одному из понятий валидности определено, что «количественной мерой критериальной валидности служат коэффициенты ранговой и бисериальной корреляции между показателями теста и критериальной мерой, задаваемой при конструировании теста» [47]. Данное определение 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 результатов 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

адекватной математической модели валидности тестов для оценки знаний и умений, а все 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. Даже применение 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 менее сложно и неоднозначно интерпретируется термин «надежность теста». Практически все приводимые в педагогической литературе определения надежности теста объединяет такое свойство теста, как воспроизводимость его результатов, 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» [25]. 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 раздельных частей теста;

– эквивалентную надежность, основанную на сравнении результатов решения испытуемым теста и его альтернативного варианта [48].

Таким образом, надежность теста может характеризоваться сходимостью последовательности его результатов. Однако хорошей воспроизводимостью (сходимостью) могут обладать и недостоверные результаты, обусловленные, например, систематической ошибкой теста. В связи с этим ряд определений надежности теста содержит упоминание о его точности. Так, некоторые ученые [49] утверждают, что «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». Н.Ф. Ефремова, напро-

тив, отмечает, что точность и устойчивость результатов тестирования оказываются вторичными по отношению к надежности: «Надежность теста – показатель точности и устойчивости результатов измерения с помощью теста при его многократном применении. Характеризует степень адекватности отражения тестом соответствующей генеральной совокупности заданий» [1].

В.С. Аванесов отмечает, 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 [50]. В качестве меры надежности применяется коэффициент корреляции. В частности, значения коэффициента корреляции выше 0,5 свидетельствуют об удовлетворительной надежности, но в практике тесты с надежностью 0,8 и менее, как правило, не применяются.

Очевидным недостатком корреляционных методов оценки надежности теста является то, что при рассмотрении тест-ретестовой надежности оценки тестируемых при каждом применении теста могут отличаться, поэтому встает вопрос о корректности выбора множеств оценок для расчета коэффициента корреляции, а также насколько корректно сопоставление результатов, если в повторных тестах полностью или частично повторяются уже известные испытуемым вопросы, а время тестирования не меняется. В других интерпретациях надежности (разделенная, эквивалентная) для обеспечения корректности результатов корреляционного анализа исходных и повторных данных требуется безукоризненное обоснование тождественности разделенных частей теста или первого и альтернативного тестов. Кроме того, корреляционный подход к расчету показателей надежности теста не решает проблему априорной оценки достоверности.

Д.И. Попов и Е.Д. Попова указывают другую меру точности и, соответственно, надежности теста: «надежность связана с понятием стандартной ошибки измерения: чем выше надежность, тем меньше стандартная ошибка измерения» [49]. Однако без указания величины, относительно которой вычисляется стандартная ошибка, при использовании типовых методик обработки статистических данных, в которых стандартная ошибка рассчи-

тывается относительно среднего значения, мера точности оказывается тождественной мере воспроизводимости.

Обратим внимание на то, что стандартная ошибка (среднеквадратическое отклонение) является квадратным корнем из дисперсии, которая, в свою очередь, характеризует корреляцию результатов тестирования, т. е. валидность теста. Таким образом, валидность и надежность теста количественно оцениваются тождественными характеристиками, что по факту так же делает их либо тождественными друг другу, либо соотносит их как общее и частное.

Дисперсия оценок характеризует и такое качество теста, как его дифференцирующая способность, т. е. способность формировать и интерпретировать оценки в широком диапазоне значений, что наряду с правильно подобранными критериями и шкалами позволяет корректно дифференцированно оценивать знания и умения обучающихся. Дифференцирующая способность подразумевает включение в тест большого количества заданий и увеличение в нем доли легких и трудных заданий, что соответствует и целям повышения содержательной валидности, обеспечивающей оценивание большого числа параметров, в том числе латентных [51, 52], однако при рассмотрении надежности исключительно как меры воспроизводимости результатов тестирования в одинаковых по уровню освоения учебного материала выборках тестируемых при соблюдении прочих тождественных условий делается парадоксальный вывод о снижении надежности теста при повышении его дифференцирующей способности и валидности, а оптимальность надежности и валидности теста может быть достигнута путем подбора заданий сбалансированной сложности [50]. Такой вывод выглядит особенно странно в контексте того, что валидность и надежность рассматриваются как основные свойства теста, характеризующие его качество и адекватность, проявляющиеся в достоверности получаемых при его применении оценок учебных достижений обучающихся, а ряд исследователей среди количественных показателей надежности и валидности указывают близкие по смыслу величины. По нашему мнению, нет оснований утверждать, что повышение надежности и обеспечение валидности теста не могут осуществляться независимо друг от друга, в связи с чем применение термина «оптималь-

ность» к надежности в принципе некорректно, а более приемлемым является термин «приемлемая надежность», или «удовлетворительная надежность».

Большего внимания заслуживает утверждение о том, что чем выше погрешности оценок теста, количественно характеризующие надежность, тем ниже тестовая эффективность [34, 53].

По нашему мнению, причиной терминологических противоречий в области оценки качества тестов является использование (в том числе вторичное заимствование, например из психологии) педагогической квалиметрией классической терминологии технических, естественно-научных и математических дисциплин с собственной трактовкой терминов на основе общих знаний или лингвистических подходов. Вторым по значимости фактором терминологических противоречий, на наш взгляд, является буквальный перенос понятий и определений теоретической тестологии в практику контроля и измерения параметров учебных достижений обучающихся без учета специфики оцениваемых объектов, что породило несогласованность теории и практики в организации проверки знаний и умений обучающихся и интерпретации ее результатов. Немаловажной проблемой методологического характера является несогласованное и подчас некорректное развитие терминологической базы в области оценки качества тестов в современных педагогических исследованиях, выражающееся в наделении широко употребляемых, устоявшихся понятий новыми признаками и свойствами, несистемное введение дополнительных частных терминов, поглощаемых по смыслу базовыми понятиями, а также терминологические новации, противоречащие понятийному аппарату смежных наук. Все это приводит к множественности и некорректности трактовок базовых терминов, подмене понятий и разрушению единого научного языкового пространства в целом и единой научной терминологии в частности.

Таким образом, существует проблема корректной оценки качества тестов для контроля знаний как в контексте терминологии, так и с точки зрения математических моделей. Отсутствие корректных математических моделей КТОЗУО сказывается на возможности априорных оценок качества тестов и точности прогнозирования результатов тестирования. До некоторой степени прогноз успешности решения теста может быть получен на осно-

ве метода Раша [35, 52, 54–56] как вероятность решения заданий определенной сложности обучающимся, если известны результаты решения данного теста (А) группой других обучающихся и относительный уровень подготовки испытуемого, определяемый по результатам решения им тестов (В) аналогичной трудности на основе других баз заданий (например, из других предметных или тематических областей). Применение данного метода дает наиболее достоверные результаты для единичных тестовых заданий – ЕТЗ (введем этот термин, обозначающий *элементарный отдельно оцениваемый компонент базы заданий теста*, для более четкого, по сравнению с популярным в педагогической литературе термином «тестовое задание», разграничения с понятием «тест») или тестов, составленных из ЕТЗ примерно одинаковой трудности, т. е. обладающих минимальной дифференцирующей способностью. При этом на откуп экспертным оценкам остается определение уровня трудностей тестов А и В, а определение уровня подготовки тестируемых требует большого объема эмпирических данных, достаточного для формирования репрезентативных выборок, или дополнительных экспертных оценок, что в конечном результате может привести к большой дисперсии оценок искомых величин, являющейся мерой их достоверности.

В значительной мере корректность прогноза успешности решения теста и достоверности получаемых на его основе оценок определяется корректностью моделей тестируемых. До настоящего времени наиболее популярные модели тестируемого представляются некоторым интегральным уровнем подготовки, который с убывающей вероятностью позволяет решать задания возрастающей трудности [13, 35, 61, 55, 56]. На этом основаны и методики создания адаптивных тестов, и применение метода Раша для согласования уровней сложности тестов и прогнозирования результатов тестирования. Такие модели не учитывают содержательные лакуны в знаниях испытуемых, в особенности на определенном этапе обучения, перекосы в соотношении «знания – умения» (например, хорошее знание теории и неумение решать задачи некоторой группой обучающихся или, наоборот, хорошие навыки решения задач при теоретической «запущенности» у другой группы обучающихся из той же выборки тестируемых) и другие особенности сформированности знаний и умений, обуслов-

ленные как индивидуальными особенностями обучающихся (уровнем базовой подготовки, способностями к тому или иному виду учебной деятельности), так и результативностью педагогической деятельности преподавателей. Учет данных нюансов 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 рассматриваемых моделей узким кругом теоретических и прикладных задач.

2.2. Обоснование применимости методологии и математического аппарата теории надежности к оцениванию качества теста и достоверности результатов тестирования

Основной идеей предлагаемого нами подхода к описанию качества ТОУД является вывод о тождественности алгоритма оценивания учебных достижений обучающихся системами тестирования алгоритмам работы автоматизированных измерительных систем (АИС), широко применяемых в технических системах контроля качества, управления объектами и технологическими процессами, безопасности, телеметрии и др. Для корректности данного вывода будем полагать, что тестирование знаний и умений обучающихся выполняется компьютерной программой, исключающей или минимизирующей, по аналогии с АИС, влияние человеческого фактора на принятие решений. Данный подход призван обосновать применимость для анализа качества ТОУД методов математического моделирования функционирования и оценки качества их технических аналогов, а также методологии и математического аппарата теории надежности технических систем и устройств.

Выбор технического объекта – аналога ТОУД – проведен нами методом экспертного опроса. На первом этапе экспертами были выдвинуты для дальнейшего анализа технические объекты, наиболее, по их мнению, сходные по принципу действия с компьютерными тестами КТОЗУО, и проведено обоснование выбора на основе критерия:

$$\sum_{j=1}^{n} p_j \cdot v_j > n_{min}, \quad (2.9)$$

где p_j – выдвигаемый экспертом верифицируемый признак сходства принципов работы технического объекта и КТОЗУО, v_j – показатель верификации, равный единице при совпадении мнения эксперта и остальных членов экспертной группы и равный нулю при несовпадении мнений экспертов, n_{min} – некоторое минимально необходимое количество согласованных экспертами признаков, определяемое по окончании процедур согласования для всех рассматриваемых технических объектов. На втором этапе по аналогичной методике проводился отбор и обоснование дополнительных преимуществ устройств (систем), модели воспроизводства которых могут быть внесены в алгоритмы формирования баллов и оценки качества тестов для КТОУЗО.

Наибольшее количество признаков сходства принципов работы с тестами для КТОУЗО и дополнительных потенциально востребованных преимуществ (выделены курсивом) выявлено у следующих автоматизированных технических устройств и систем.

1. Электронные весы – измерение с преобразованием первичного информационного сигнала в необходимый для обработки АИС формат; отображение результатов измерения в удобной для восприятия и первичного анализа форме; возможность сравнения результатов измерения с эталоном; возможность калибровки (настройки); *возможность учета систематической погрешности и внесения поправок*; *возможность проведения многократных повторных измерений для неизменных объектов и условий измерений; нормированная погрешность как мера точности измерений*.

2. АИС контроля состояния объекта (технологического процесса) – адаптация к параметрам контролируемого объекта; автоматизированный анализ измерительной информации; информирование о важнейших параметрах состояния объекта; применение алгоритмов диагностики; применение алгоритмов сравнения с идеальным состоянием; учет допустимых отклонений от идеальных параметров объекта; принятие решения о соответствии или несоответствии объекта требуемым параметрам; *возможность*

многопараметрических измерений; возможность диалогового режима с оператором.

3. Система контроля управления доступом (СКУД) – автоматизированный анализ измерительной информации; возможность сравнения результатов измерения с эталоном; возможность многоэтапного контроля с принятием решения на каждом этапе; применение алгоритмов сравнения с идеальным состоянием; учет допустимых отклонений от идеальных параметров объекта; принятие решения о соответствии или несоответствии объекта требуемым параметрам; автокалибровка; изменяемый массив данных для сравнения с результатами измерений; *возможность проведения многократных повторных измерений для неизменных объектов и условий измерений; возможность учета информационных помех.*

4. Сортировочный (отбраковочный) автомат – автоматизированный анализ измерительной информации; возможность сравнения результатов измерения с эталоном; возможность многоэтапного контроля с принятием решения на каждом этапе; применение алгоритмов сравнения с идеальным состоянием; применение алгоритмов сортировки; учет допустимых отклонений от идеальных параметров объекта; принятие решения о соответствии или несоответствии объекта требуемым параметрам; применение многоуровневой оценочной шкалы; возможность калибровки (настройки); *возможность проведения многократных повторных измерений для неизменных объектов и условий измерений; возможность многопараметрических измерений.*

Анализ признаков сходства приведенных технических объектов с тестом для КТОЗУО позволяет сделать следующие выводы:

– наиболее близким по принципу действия к КТОЗУО является сортировочный (отбраковочный) автомат;

– сортировочный (отбраковочный) автомат и СКУД обладают почти идентичным набором признаков сходства с КТОЗУО по принципу действия.

На основании проведенного анализа признаков сходства будем полагать, что КТОЗУО по своему предназначению и алгоритму функционирования является автоматом по отбраковке (при двухбалльной оценочной шкале «зачет-незачет») или сортировке (при многобалльной оценочной шкале) объектов по некоторому

заданному параметру. Одной из распространенных реализаций такого автомата является СКУД, которая может выполнять функции допуска-недопуска на объект или разграничения (определения уровня) доступа по вводимому паролю (цифровому коду) или биометрическим данным.

Как в СКУД или ином сортировочном автомате, так и в КТОЗУО можно выделить следующие структурные элементы:

– блок ввода информации (интерфейс);

– измерительный блок (в КТОЗУО роль измерительных преобразователей выполняют ЕТЗ, преобразующие вводимую информацию в бинарный код («верно-неверно») или иную числовую или символьную форму, удобную для обработки и визуального восприятия);

– блок генерации запросов (формирования комплектов заданий из базы);

– библиотека ключей (паролей, правильных ответов);

– блок сравнения;

– блок принятия решения на основе заданного критерия, включающего в себя пороговые значения (одно или несколько) и решающие правила.

Все блоки КТОЗУО, кроме первого, могут быть реализованы как в автоматическом режиме (в компьютерной программе, специализированном электронном устройстве), так и непосредственно специалистом, проводящим тестирование.

Данное сходство позволяет рассмотреть применимость в моделях оценки качества КТОЗУО элементов теории надежности технических систем.

ГОСТ 27.002-89 определяет надежность как «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.

Надежность технических систем (изделий) оценивают по четырем показателям: безотказность, долговечность, ремонтопригодность и сохраняемость.

Безотказность – свойство объекта выполнять заданные функции непрерывно в течение некоторого времени или некоторой наработки. Применительно к КТОЗУО можно сказать, что безотказность – свойство безошибочно соотносить знания испытуемых с принятой оценочной шкалой в течение некоторого времени или (и) некоторого количества тестирований. Как и для технических устройств, на отказ КТОЗУО, т. е. утрату способности правильно оценивать знания тестируемых, оказывает влияние не только количество применений (рабочих циклов), но и их интенсивность, т. е. зависимость от времени эксплуатации может носить неоднозначный характер.

Долговечность – свойство объекта выполнять заданные функции при установленной системе поддержания его работоспособности до наступления предельного состояния, при котором дальнейшая эксплуатация объекта недопустима или нецелесообразна. Применительно к КТОЗУО поддержание его работоспособности может осуществляться путем изменения объема и содержания базы заданий и базы знаний, порогов и диапазонов оценок, т. е. залогом долговечности КТОЗУО является его регулярное обновление. Долговечность является комплексным параметром, зависящим от ремонтопригодности и сохраняемости объекта.

Ремонтопригодность 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. Ремонтопригодность определяется степенью доступности функций настройки параметров алгоритма оценивания выполнения тестовых заданий, подсчета баллов и формирования общих оценок, а также базы заданий и базы знаний для оператора или администратора тестовой системы. Она может измеряться в относительных единицах [процентах или в пределах от нуля (неремотопригодность) до единицы (полная ремонтопригодность)] на основе как экспертных оценок, так и оценочных алгоритмов, формали-

зующих значимость и степень изменения отдельных элементов теста. Сохраняемость измеряется в единицах времени и может определяться как априорно (экспертными оценками), так и апостериорно (по результатам тестирования или экспертным оценкам).

Очевидно, что из четырех приведенных характеристик лишь безотказность может оказаться пригодной для количественных оценок надежности КТОЗУО.

Одной из частных реализаций КТОЗУО являются компьютерные контролирующие программы. Безотказность является одной из четырех составляющих функциональной надежности программных продуктов наряду с работоспособностью, безопасностью и защищенностью [57]. Таким образом, безотказность представляет собой универсальный показатель надежности КТОЗУО, общий для его различных реализаций (аппаратных, программных, программно-аппаратных), а также, с учетом применимости к текстовым документам, и ТОУД в целом.

Исходя из положений теории надежности [47, 58], безотказность КТОЗУО по аналогии с техническими средствами определяется вероятностью его работы без отказа в течение определенного периода времени.

Отказ ТОУД, компьютерного или на бумажном носителе, как и отказ СКУД, 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. Очевидно, для всего теста ошибки первого рода будут приводить к занижению оценок знаний испытуемых.

Причинами ошибок первого рода могут быть:

1) ошибка отбора содержания теста (включение в него неизученных вопросов или вопросов из других тем, разделов, предметных областей, плохо известных обучающимся, т. е. то, что может быть и причиной низкой содержательной валидности);

2) некорректная формулировка вопроса и (или) ответов, приводящая к неоднозначному пониманию их смысла тестируемым;

3) ошибка сопоставления вопроса и ответа, когда правильный ответ считается неправильным или правильный ответ отсут-

ствует (из-за невнимательности или низкой квалификации составителя);

4) ошибка программирования теста (для КТОЗУО);

5) ошибка программной оболочки теста (для КТОЗУО);

6) ошибка ввода ответа тестируемым (например, из-за невнимательности).

Ошибка второго рода для ответа на один вопрос проявляется в том, что испытуемый не знает правильный ответ, но система расценивает его ответ как правильный. Ошибки второго рода приводят к завышению оценок знаний 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, равно как и причины низкой содержательной валидности);

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 программирования теста (для КТОЗУО);

5) ошибка программной оболочки теста (для КТОЗУО);

6) угадывание тестируемым правильного ответа;

7) запоминание тестируемым правильного ответа без знания (из опыта прошлого тестирования);

8) подсказка тестируемому правильного ответа в ходе или накануне проведения теста;

9) использование тестируемым разного рода шпаргалок.

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 резуль-

тат. 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.

По аналогии с техническими системами по отношению к КТОЗУО можно интерпретировать классификацию отказов по ГОСТ 27.002-89. Так, по возможности дальнейшего использования или восстановления объекта отказы разделяют на *полные* и *частичные*. Применительно к КТОЗУО частичный 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., т. е. содержанию обучения (ошибка второго рода), что особенно характерно для общественно-научных и юридических учебных дисциплин. Характерные примеры: принятие поправок в Конституцию Российской Федерации 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 в процедуре начисления баллов и задании оценочной шкалы или порога положительной оценки.

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 обучающегося или проявившуюся для большинства тестируемых в группе. Помимо собственно недостоверной оценки или суммы баллов, частичные и полные отказы в соответствующей мере 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 педагогиче-

ских экспериментов и др. Немаловажным фактором возникновения частичных отказов КТОЗУО при групповом тестировании, в особенности при проверке остаточных знаний [59, 60], играет низкая мотивация отдельных обучающихся.

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 первого или, чаще, второго рода, и проявляется в постепенном (от одной группы тестируемых к другой) увеличении дисперсии отклонения (уменьшении коэффициента корреляции) 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 несоответствие уровня заданий уровню остаточных знаний (слишком легкие или слишком трудные ЕТЗ, приводящие к ошибкам соответственно второго и первого рода), несогласованность содержания тестов со структурой остаточных знаний (низкая содержательная валидность, обусловливающая возникновение ошибок первого рода), а также предварительное ознакомление обучающихся с заданиями теста и (или) ответами на них.

Очевидные трактовки применительно к КТОЗУО могут обосновать проявление в них таких характерных для технических объектов видов отказов, как *устранимые* и *неустранимые*, *систематические, катастрофические* и др.

Количественными характеристиками отказов являются вероятность отказа P_0 и интенсивность отказов $\lambda_{\text{отк}}(t)$.

Интенсивность отказов является статистическим показателем и применительно к КТОЗУО может определяться как отношение количества неверных оценок, полученных в результате тестирования знаний обучающихся, к общему количеству таких оценок за некоторый период времени. График $\lambda_{\text{отк}}(t)$ представ-

ляется корытообразной кривой [47, 58], на которой выделяют три характерных участка (рис. 2.1).

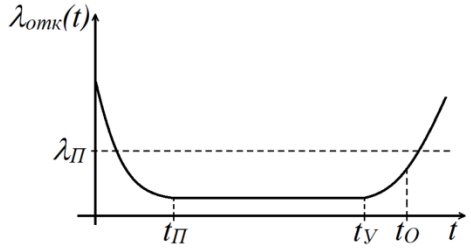


Рис. 2.1. Типичный график зависимости интенсивности отказов от времени

Первый участок приходится на временной интервал [0; $t_п$], где $t_п$ – 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 значение $\lambda_{\text{отк}}(t)$ на втором интервале [$t_п$; $t_у$], где $t_у$ – время устаревания теста, обусловлены наличием неустраненных ошибок КТОЗУО и случайными причинами возникновения ошибок первого и второго рода. На третьем временном интервале [$t_у$; $t_о$], где $t_о$ – 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 дидактических единиц учебной дисциплины, неуклонно увеличивается от тестирования к тестированию. Отсутствие заинтересованности разработчика по корректировке 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.

Следует отметить, что 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. Так же на графике $\lambda_{\text{отк}}(t)$ может отсутствовать 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, будет равна сумме вероятностей ошибок первого и второго рода:

$$P_O = P_{O1} + P_{O2.} \tag{2.10}$$

При эксплуатации теста, в основном за период времени [0; $t_{\text{п}}$], разработчики выявляют и исправляют систематические ошибки первого рода 1–5 и частично ошибки второго рода 1–5, а также в течение всего периода эксплуатации принимают меры по устранению ошибок второго рода 8 и 9).

Снижение вероятности ошибок первого и второго рода:

$$P'_{01} = P'_{\text{нач1}} \cdot e^{\frac{-t \cdot \alpha_{k1}}{\Delta t_k}}, \tag{2.11}$$

$$P'_{02} = P'_{\text{нач2}} \cdot e^{\frac{-t \cdot \alpha_{k2}}{\Delta t_k}}, \tag{2.12}$$

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 появления ошибок второго рода 6 и 7 интенсивность отказов теста возрастает. Соответственно, вероятность ошибок второго рода, обусловленная данными факторами:

$$P''_{02} = P''_{\text{нач3}} \cdot e^{\gamma t}, \tag{2.13}$$

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

Несложно заметить, что сумма убывающих (2.11), (2.12) и возрастающей (2.13) экспонент дает нам корытообразную кривую (рис. 2.1).

Таким образом, проблема корректной оценки качества тестов для оценивания учебных достижений обучающихся требует применения универсального подхода, способного устранить имеющиеся противоречия в терминологии, моделях и интерпретации результатов тестирования. На основании сходства по принципу действия КТОЗУО с автоматом по отбраковке или сортировке объектов по некоторому заданному параметру в оценках

качества КТОЗУО возможно и целесообразно применение элементов теории надежности технических систем – классической отрасли знания с устойчивыми методологическими основами и хорошо разработанным математическим аппаратом. 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 сложности теста, динамических воздействий и др.

3. Математические модели и оценка надежности тестов для проверки знаний и умений обучающихся

3.1. Математическая модель оценки и контроля безотказности теста из заданий с выбором единственно верного ответа

Одна из проблем, с которыми сталкиваются разработчики ТОУД, – определение оптимального (минимально достаточного) количества вопросов, обеспечивающих объективность оценки уровня подготовки обучающегося, от которого напрямую зависит трудоемкость разработки. Данная проблема актуализировалась в связи с необходимостью формирования фондов оценочных средств для проверки компетенций обучающихся и выпускников образовательных организаций высшего образования в соответствии с требованиями федеральных государственных образовательных стандартов 3-го поколения (ФГОС-3) и их модификаций (ФГОС-3+ и ФГОС-3++). Корректность решения задачи определения количества вопросов в тесте и в базе заданий в целом в значительной степени зависит от корректности математической модели надежности теста.

В процессе многократного применения КТОЗУО систематически встает вопрос о достоверности результатов тестирования, вызванной адаптацией обучающихся к заданиям, наиболее тривиально проявляющейся в заучивании ими правильных ответов без усвоения содержания изучаемого материала.

В построении математической модели надежности КТОЗУО исходим из отождествления алгоритмов функционирования КТОЗУО и браковочного (при двухбалльной оценочной шкале «зачет-незачет») или сортировочного (при многобалльной оценочной шкале) автомата объектов по некоторому заданному параметру, что позволяет использовать для оценок достоверности результатов тестирования понятийный и математический аппараты теории надежности [61].

Очевидно, по аналогии с надежностью технических устройств, надежность ТОУД означает его способность в течение

заданного времени правильно (с допустимой погрешностью) соотносить знания испытуемых с принятой оценочной шкалой.

Важной характеристикой надежности является безотказность – свойство объекта сохранять работоспособное состояние непрерывно в течение некоторого времени или некоторой наработки [62]. Применительно к ТОУД можно сказать, что безотказность – свойство ТОУД безошибочно соотносить знания испытуемых с принятой оценочной шкалой в течение некоторого времени или (и) некоторого количества тестирований. Если оценка, полученная обучающимся в результате тестирования, не соответствует его реальным знаниям (например, не согласуется с достоверным результатом других проверок), это явление должно расцениваться как отказ теста.

Отказ ТОУД проявляется в ошибках первого и второго рода [61]. Для одного вопроса теста ошибкой первого рода является незачет правильного ответа, квалификация его как неправильного. Для ТОУД в целом ошибки первого рода будут приводить к занижению оценок знаний. Ошибка второго рода для одного вопроса теста в узком смысле проявляется в расценивании неправильного ответа как правильного, а в широком смысле – в квалификации ответа как правильного при реальном незнании соответствующего учебного материала, например в результате подсказки или угадывания. Ошибки второго рода приводят к завышению оценок знаний обучающихся по итогам тестирования.

Как и для технических устройств, на отказ ТОУД, т. е. утрату способности правильно оценивать знания тестируемых, оказывает влияние не только количество применений (рабочих циклов), но и их интенсивность, т. е. зависимость от времени эксплуатации может носить неоднозначный характер.

Количественными характеристиками отказов являются вероятность отказа $P_О$ и интенсивность отказов $\lambda_{отк}(t)$ [47, 57].

Вероятность безотказной работы теста с учетом (2.10):

$$P_Б = 1 - P_О = 1 - (P_{О1} + P_{О2}), \qquad (3.1)$$

Угадывание правильного ответа при его незнании – наиболее часто встречающаяся ситуация при проверке знаний с помощью тестов с выбором верного ответа из предлагаемого набора, приводящая к ошибкам второго рода.

Вероятность угадывания ответа на вопрос в единичном тестовом задании (ЕТЗ) закрытого типа с выбором единственно верного ответа:

$$P_{\text{угад}} = \frac{1}{n - n_{\text{искл}}}, \quad (3.2)$$

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 ЕТЗ с выбором единственно верного ответа ошибкам второго рода за счет успешных попыток угадывания испытуемым правильных ответов при их реальном незнании приводит к значительным ошибкам второго рода в целом КТОЗУО, составленных из ЕТЗ данного типа, и существенно повышает вероятность их отказа (незаслуженных положительных оценок по шкале «зачет-незачет» или завышенных оценок при их многоуровневой дифференцированной шкале). Поскольку тесты, сформированные из ЕТЗ с выбором единственно верного ответа, являются наиболее распространенными среди КТОЗУО, скептическое отношение многих педагогов-практиков к достоверности оценок знаний и умений обучающихся, полученных на основе данной разновидности тестов, распространяется и на КТОЗУО в целом. Повышение вероятности безотказной работы тестов рассматриваемого типа достигается, как правило, за счет регулярного изменения базы ЕТЗ, из которой производится выборка заданий при формировании теста, или редактирования самих ЕТЗ (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 и применением ЕТЗ с усложненной структурой (например, с множественным выбором верных ответов, установлением соответствий и др. [63]), а также усложнением процедур проверки результатов тестирования (например, на основе двухступенчатых алгоритмов, адаптивных алгоритмов и других реализаций обратных связей в процедурах подбора ЕТЗ в режиме реального времени и начисления баллов). В ряде случаев, как, например, в тестах с множественным выбором, снижение вероятности ошибки второго рода приводит к увеличению вероятности

ошибки первого рода или применению эвристических критериев сомнительной корректности.

Величина $n_{\text{искл}}$ является функцией времени с некоторой характерной для данного процесса и 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:

$$n_{\text{искл}}(t) = n^{*}_{\text{искл}} + n^{\text{обуч}}_{\text{искл}}(t) - n^{\text{забыв}}_{\text{искл}}(t), \qquad (3.3)$$

где $n^{*}_{\text{искл}} \geqslant 0$ – некоторый незабываемый остаток.

Функция, характеризующая накопление и восстановление знаний, по нашему мнению, должна представлять собой сумму парциальных ступенчатых функций, величина которых $n_i^{\text{обуч}}$ отражает объем знаний, соотнесенный с ДЕСО, а времена «включения» в общем случае связаны функциональной зависимостью с $\tau_{\text{искл}}$: $t_i = f(\tau_{\text{искл}})$ – интенсивность его изучения и восстановления:

$$n^{\text{обуч}}_{\text{искл}}(t) = \sum_{i=1}^{v(t_{\text{обуч}})} n_i^{\text{обуч}} \sigma(t - t_i), \qquad (3.4)$$

где $\sigma(t - t_i)$ – функция Хевисайда [64], а предел суммирования $v(t_{\text{обуч}})$ определяется временем обучения (подготовки к тестированию) $t_{\text{обуч}}$. Функция $f(\tau_{\text{искл}})$ определяет зависимость моментов времени, характерных для приращения знаний, от временных параметров организации образовательного процесса и отражает цикличность, соотнесенную с $\tau_{\text{искл}}$, учебных занятий и самостоятельной работы обучающихся, в том числе в процессе подготовки к различным видам контроля знаний.

Для моделирования забывания больше подходит экспоненциальная функция:

$$n^{\text{забыв}}_{\text{искл}}(t) = \lceil n_0 \cdot e^{-\beta t} \rceil, \qquad (3.5)$$

где β – показатель забывчивости, определяемый индивидуальными особенностями тестируемого (памятью, ассоциативным мышлением, владением методиками удержания в памяти информа-

ции) и системностью рассматриваемых знаний, ⌈x⌉ – округление числа x в большую сторону.

С учетом выражений (3.2), (3.4), (3.5):

$$P_{\text{угад}}(t) = \left[n - n^{*}_{\text{искл}} - \sum_{i=1}^{v(t_{\text{обуч}})} n_i^{\text{обуч}} \sigma(t - t_i) + \lceil n_0 \cdot e^{-\beta t} \rceil \right]^{-1}. \quad (3.5\text{а})$$

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

$$M_{\text{пр}}(t) = M_{\text{зн}}(t) + \Delta M_{\text{уг}}(t). \quad (3.6)$$

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

$$\frac{M_{\text{пр}}}{M_{\text{зач}}} \geq 1, \quad (3.7)$$

или

$$\frac{M_{\text{зн}}(t) + \Delta M_{\text{уг}}(t)}{M_{\text{зач}}} \geq 1. \quad (3.7\text{а})$$

Отсюда необходимое условие безотказной работы теста:

$$\frac{M_{\text{зн}}(t)}{M_{\text{зач}}} \geq 1. \quad (3.7\text{б})$$

Вероятность угадывания $M_{\text{зач}}$ ответов из N вопросов теста вычисляется по формуле Бернулли [46]:

$$P_N(M_{\text{зач}}) = C_N^{M_{\text{зач}}} \cdot P_{\text{угад}}^{M_{\text{зач}}} \cdot q^{N - M_{\text{зач}}}, \quad (3.8)$$

где $P_{\text{угад}}$ – вероятность угадывания ответа на один вопрос;

$q = 1 - P_{\text{уг}}$;

$C_N^{M_{\text{зач}}} = \frac{N!}{M_{\text{зач}}! \cdot (N - M_{\text{зач}})!}$.

Если тестируемый не выполняет условие (3.7б), то для получения положительной оценки он должен угадать хотя бы $\Delta M_{\text{уг}}(t) = M_{\text{зач}} - M_{\text{зн}}(t)$ ответов. Вероятность этого события:

$$P_{N - M\text{зн}}(M_{\text{зач}} - M_{\text{зн}}) = C_{N - M\text{зн}}^{M_{\text{зач}} - M_{\text{зн}}} \cdot P_{\text{угад}}^{M_{\text{зач}} - M_{\text{зн}}} \cdot q^{N - M\text{зн} - M_{\text{зач}}}. \quad (3.9)$$

Будем полагать, что в базе находятся W вопросов из разных тем (индексы соответствуют номерам тем): $W = W_1 + W_2 + W_3 + \ldots$ Поскольку сложность вопросов различается, условно разделим ее на три уровня: легкий, средний и трудный, т. е. для каждой темы:

$$W_i = W_{i\,\text{легк}} + W_{i\,\text{ср}} + W_{i\,\text{тр}}. \quad (3.10)$$

Если каждая i-я тема содержит d_i ДЕСО, то количество вопросов, которое ей будет соответствовать:

$$W_i = \sum_{j=1}^{d_i} D_{i,j}. \quad (3.11)$$

Здесь j – номер ДЕСО в i-й теме, $D_{i,j}$ – количество вопросов для j-й ДЕ. Тогда общее количество вопросов:

$$W = \sum_{i=1}^{k} W_i = \sum_{i=1}^{k} \sum_{j=1}^{d_i} D_{i,j}\ , \quad (3.12)$$

где k – число тем в дисциплине.

С учетом уровней сложности вопросов:

$$D_{i,j} = W_{i,j}^{\text{легк}} + W_{i,j}^{\text{ср}} + W_{i,j}^{\text{тр}} = \sum_{n=1}^{n_{\text{легк}}} W_{i,j,n}^{\text{легк}} + \sum_{n=1}^{n_{\text{ср}}} W_{i,j,n}^{\text{ср}} + \sum_{n=1}^{n_{\text{тр}}} W_{i,j,n}^{\text{тр}}. \quad (3.13)$$

Разделение вопросов по уровням сложности позволяет вводить дополнительные критерии удовлетворительной оценки, т. е. повышать надежность теста за счет уменьшения вероятности ошибок второго рода. Присвоим каждому уровню сложности весовой коэффициент: $P_{\text{легк}}$, $P_{\text{ср}}$, $P_{\text{тр}}$. Тогда условием успешного прохождения теста будет набор заданного количества $T_{\text{зач}}\,min$ баллов:

$$N_{\text{легк}}^{\text{отв}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{отв}} \cdot P_{\text{ср}} + N_{\text{тр}}^{\text{отв}} \cdot P_{\text{тр}} \geq T_{\text{зач}}\,min, \quad (3.14)$$

где $N_{\text{легк}}^{\text{отв}}$, $N_{\text{ср}}^{\text{отв}}$, $N_{\text{тр}}^{\text{отв}}$ – количество правильных ответов на вопросы, отнесенные соответственно к категориям легких, средней трудности и трудных.

Нормальной для тестирования является ситуация, когда посредственно подготовленный испытуемый не отвечает на сложные 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 сложные вопросы больше, чем процент безошибочно отвеченных легких вопросов, свидетельствует скорее о «взломе» теста или использовании испытуемым подсказок. Таким образом, разделе-

ние вопросов по уровням сложности позволяет контролировать корректность процедур тестирования и надежность самих тестов.

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

$$N_{\text{легк}}^{*\,\text{отв}} \cdot P_{\text{легк}} + N_{\text{ср}}^{*\,\text{отв}} \cdot P_{\text{ср}} \geq T_{\text{зач}\,min}; \qquad (3.14\text{а})$$

$$N_{\text{легк}}^{*\,\text{отв}} \geq N_{\text{легк}}^{\text{отв}};\ N_{\text{ср}}^{*\,\text{отв}} \geq N_{\text{ср}}^{\text{отв}}.$$

Очевидно, правильные ответы на все легкие и средней трудности вопросы должны гарантировать положительную оценку, т. е.:

$$N_{\text{легк}}^{\text{вопр}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{вопр}} \cdot P_{\text{ср}} > T_{\text{зач}\,min}. \qquad (3.15)$$

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

$$N_{\text{легк}}^{\text{вопр}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{вопр}} \cdot P_{\text{ср}} = \frac{T_{\text{зач}\,min}}{k^*}. \qquad (3.16)$$

Здесь $\mathrm{k}^* < 1$ – 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:

$$N_{\text{легк}}^{\text{отв}} \geq N_{\text{легк}\,min}^{\text{отв}} = N_{\text{легк}}^{\text{вопр}} \times k_{\text{легк}}^{*}; \qquad (3.17\text{а})$$

$$N_{\text{ср}}^{\text{отв}} \geq N_{\text{ср}\,min}^{\text{отв}} = N_{\text{ср}}^{\text{вопр}} \times k_{\text{ср}}^{*}. \qquad (3.17\text{б})$$

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

Пример 1. Тест содержит 50 вопросов: 30 легких, 15 средней трудности и 5 трудных. Определим весовые коэффициенты уровней сложности следующим образом: $P_{\text{легк}} = 0{,}5$, $P_{\text{ср}} = 1$, $P_{\text{тр}} = 1{,}5$. Будем полагать, что для успешного прохождения теста достаточно правильно ответить на 80 % легких и средней трудности вопросов ($k_{\text{легк}}^{*} = k_{\text{ср}}^{*} = 0{,}8$). Тогда

$$T_{\text{зач}\,min} = (30 * 0{,}5 + 15 * 1) * 0{,}8 = 24 \text{ балла};$$

Если $k_{\text{легк}}^{*} = k_{\text{ср}}^{*} = k^*$, то

$$N_{\text{легк}}^{\text{отв}} \geq N_{\text{легк}\,min}^{\text{отв}} = 30 \times 0{,}8 = 24;$$

$$N_{\text{ср}}^{\text{отв}} \geq N_{\text{ср}\,min}^{\text{отв}} = 15 \times 0{,}8 = 12,$$

т. е. $M_{\text{зач}} = N_{\text{легк}\,min}^{\text{отв}} + N_{\text{ср}\,min}^{\text{отв}} = 36.$

Однако правильным было бы предположить, что $k_{\text{легк}}^{*} > k_{\text{ср}}^{*}$, т. е. испытуемый отвечает на больший процент легких вопросов, чем вопросов средней трудности. Тогда набор количества баллов, превышающего $T_{\text{зач}\,min}$, при несоблюдении условия (3.7) может быть расценен как ошибка второго рода.

В общем случае критерий удовлетворительной оценки может быть дополнен требованием решения какого-то количества трудных вопросов ($k_{\text{тр}}^{*} > 0$).

Очевидно, во избежание противоречивых толкований результатов разработчики теста в этом случае должны соблюдать условие:

$$N_{\text{легк}}^{\text{вопр}} \cdot P_{\text{легк}} + N_{\text{ср}}^{\text{вопр}} \cdot P_{\text{ср}} < T_{\text{зач}\,min}. \qquad (3.15\text{а})$$

На рис. 3.1 представлен пример дерева событий прохождения теста с дополнительными скрытыми для тестируемых требованиями на решение определенного количества легких, средней сложности и сложных вопросов [65].

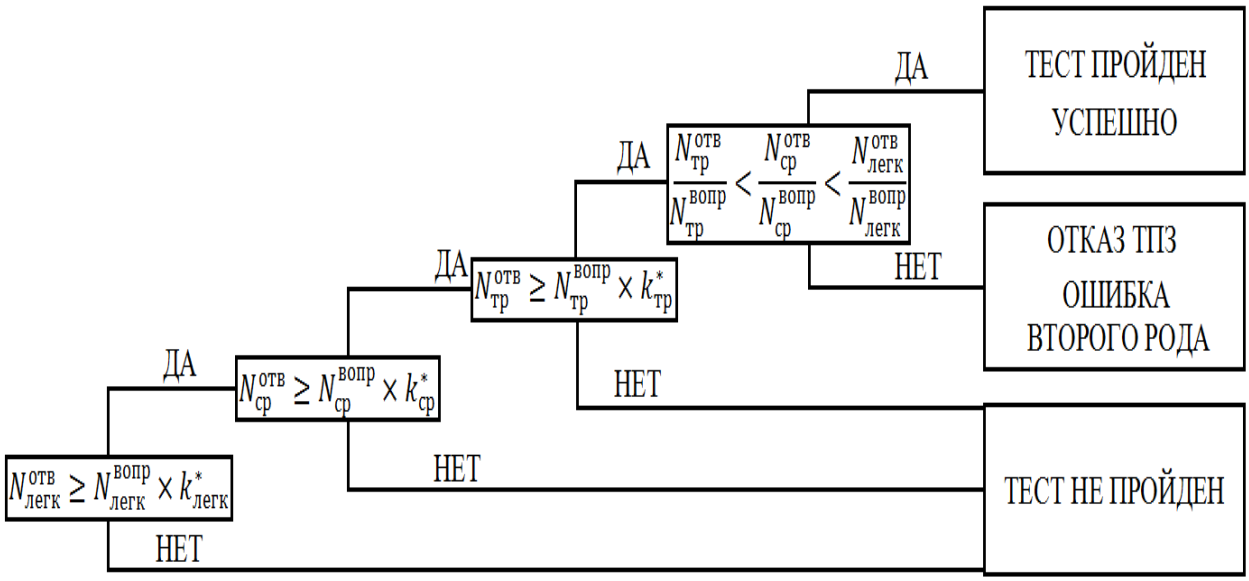


Рис. 3.1. Пример дерева событий прохождения теста

Условие (3.13) с учетом трех дополнительных требований и соблюдения требования, определяемого неравенством (3.15а), запишется:

$$N_{легк}^{отв} \cdot P_{легк} \cdot \varphi_{легк} + N_{ср}^{отв} \cdot P_{ср} \cdot \varphi_{ср} + N_{тр}^{отв} \cdot P_{тр} \cdot \varphi_{тр} \geq T_{зач\,min}, \quad (3.14\text{б})$$

где φ определяется следующим образом:

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

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

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

Пример 2. Тест состоит из 50 вопросов: 30 легких ($P_{легк} = 0{,}5$,), 15 средней трудности ($P_{ср} = 1$), 5 сложных ($P_{тр} = 1{,}5$). Условия успешного прохождения теста: необходимо дать правильные ответы на 90 % легких вопросов и 80 % вопросов средней трудности.

Отсюда $T_{зач\,min}$ =25,5 баллов; $N_{легк\,min}^{отв} = 27$; $N_{ср\,min}^{отв} = 12$.

Рассмотрим модели обучающихся, условно отнесенных к шести уровням освоения учебного материала, характеризующихся знанием определенного процента (количества) ответов на вопросы различной категории сложности. Будем полагать, что в зависимости от уровня подготовки обучающиеся могут исключить из вариантов неизвестных им ответов определенное количество неверных ответов (от нуля до двух), тем самым повысить вероятность угадывания верного ответа. Например, первый (низший) уровень подготовки характеризуется знанием 50 % легких вопросов и возможностью исключения двух неверных вариантов ответов на остальные легкие вопросы, незнанием ответов на вопросы средней трудности и возможностью исключения одного неверного варианта ответа на вопросы данной категории, незнанием ответов и невозможностью исключения их неверных вариантов для сложных вопросов. Характеристики уровней освоения учебного материала (подготовки) приведены в табл. 3.1.

Таблица 3.1

Количественные характеристики освоения учебного материала обучающимися

Уровень подготовки	Категория сложности вопроса	Процент известных вопросов	Количество известных вопросов	Количество исключаемых вариантов ответа	Вероятность угадывания
1	легкие	50	15	2	1/2
	средней трудности	0	0	1	1/3
	трудные	0	0	0	1/4
2	легкие	60	18	2	1/2
	средней трудности	20	3	1	1/3
	трудные	0	0	0	1/4
3	легкие	70	21	2	1/2
	средней трудности	50	7	2	1/2
	трудные	0	0	1	1/3
4	легкие	80	24	2	1/2
	средней трудности	60	9	2	1/2
	трудные	0	0	1	1/3
5	легкие	90	27	2	1/2
	средней трудности	70	10	2	1/2
	трудные	0	0	2	1/2
6	легкие	90	27	2	1/2
	средней трудности	90	13	2	1/2
	трудные	0	0	2	1/2

В табл. 3.2 приведены вероятности выполнения условий прохождения теста испытуемым в зависимости от уровня его подготовки. Наиболее подготовленный испытуемый (шестой уровень) знает 90 % легких вопросов (в нашем примере это дает ему 13,5 балла) и 90 % вопросов средней сложности (13 баллов), что позволяет ему набрать 26,5 балла и тем самым выполнить основное условие успешного прохождения теста, определяемое выражением (3.14), а также дополнительные условия (3.17а) и (3.17б). Таким образом, испытуемый с вероятностью P=1 успешно пройдет тест.

Таблица 3.2

Результаты моделирования прохождения теста

Уровень подготовки	Категория сложности вопроса	Набранные баллы	Выполнение условий						Вероятность прохождения теста	Средний прогнозируемый балл
			3.14б		3.17а		3.17б			
			Недостающие баллы	Вероятность выполнения	Недостающие ответы	Вероятность выполнения	Недостающие ответы	Вероятность выполнения		
1	легкие	7,5	18	0,0209	12 из 15	0,0176			$1{,}04\cdot10^{-7}$	18,125
	средней трудности	0					12 из 15	$2{,}82\cdot10^{-3}$		
2	легкие	9	13,5	0,0693	9 из 12	0,073			$1{,}93\cdot10^{-5}$	20,875
	средней трудности	3					9 из 12	$3{,}82\cdot10^{-3}$		
3	легкие	10,5	8	0,9625	6 из 9	0,254			0,088	26,25
	средней трудности	7					5 из 8	0,3633		
4	легкие	12	4,5	0,9964	3 из 6	0,656			0,4292	28,0
	средней трудности	9					3 из 6	0,656		
5	легкие	13,5	2	0,9998	0	1			0,8124	30,5
	средней трудности	10					2 из 5	0,8125		
6	легкие	13,5	0	1	0	1			1	32,0
	средней трудности	13					0	1		

Испытуемый с пятым уровнем подготовки знает 90 % легких вопросов, что дает ему 13,5 балла, и 70 % вопросов средней сложности (10 баллов). Для выполнения условия (3.14) ему не хватает 2 баллов. Эти баллы он может набрать, попытавшись угадать ответы на оставшиеся вопросы, причем его знания позволяют исключить два неверных варианта ответа (во всех категориях вопросов) и выбирать из двух оставшихся вариантов. Тогда вероятность угадывания ответа на один вопрос: $P_{угад} = \frac{1}{4-2} = \frac{1}{2}$. Испытуемый наберет недостающие баллы, если угадает ответы на вопросы в любой из следующих комбинаций:

1) 1 легкий, 1 трудный;

2) 2 легких, 1 средней трудности;

3) 1 средней трудности, 1 трудный;
4) 2 средней трудности;
5) 2 трудных.

Вероятность угадать ответы на один легкий и один трудный вопросы получим, умножив вероятности угадывания хотя бы одного вопроса из оставшихся легких (3 вопроса) и трудных (5 вопросов) соответственно. Таким образом, для приведенных выше комбинаций вероятности угадывания соответственно равны: 0,848; 0,493; 0,9385; 0,8125; 0,8125. Вероятность того, что хотя бы одна из этих комбинаций будет реализована, найдем по формуле [46]:

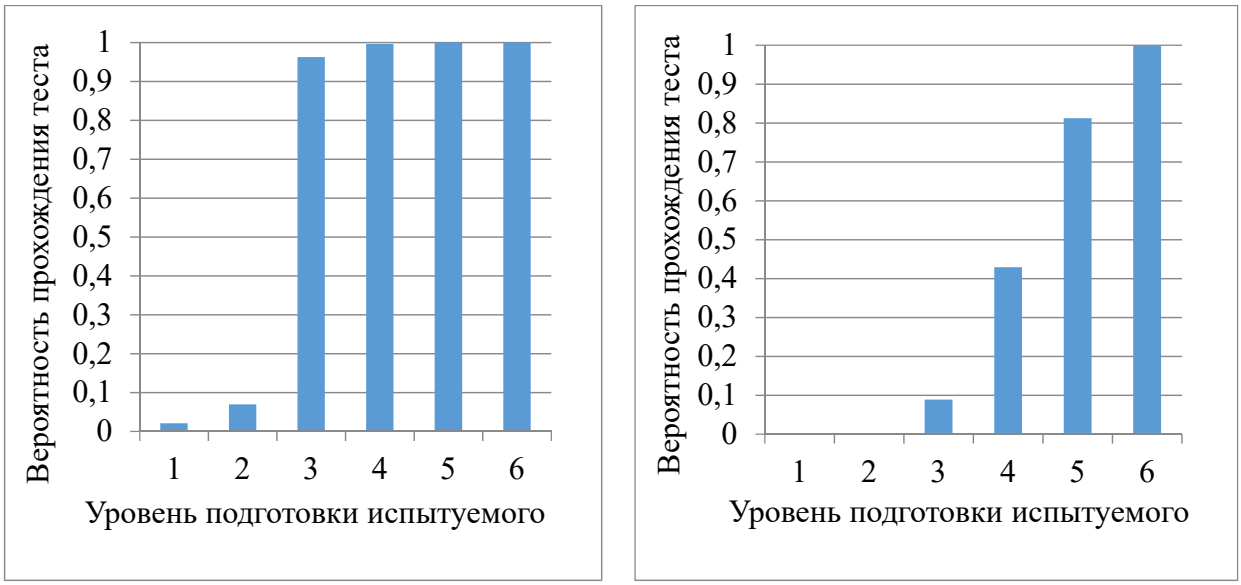
$$P = 1 - q_1 * q_2 * ... * q_n, \quad (3.18)$$

где $q_i = 1 - p_i$, n – количество комбинаций, в результате применения которой для испытуемого с пятым уровнем подготовки вероятность набрать недостающие баллы и выполнить условие (3.14) оказывается равной 0,99983. Дополнительное условие (3.17а) выполнено с вероятностью P=1. Для выполнения дополнительного условия (3.17б) испытуемому с пятым уровнем подготовки необходимо угадать ответы хотя бы на два вопроса средней сложности из оставшихся пяти. Учитывая, что вероятность угадать ответ на один вопрос $P_{угад} = \frac{1}{2}$, получим вероятность выполнения условия (3.17б) равной 0,8125. Вероятность выполнения всех трех условий получим, перемножив вероятности выполнения каждого из условий. Для рассматриваемого случая вероятность успешного прохождения ТОУД равна 0,8124.

Аналогичным способом рассчитаны вероятности прохождения ТОУД для остальных испытуемых. Число комбинаций для выполнения первого условия испытуемыми с первым уровнем подготовки составило 46, со вторым уровнем – 40, с третьим – 28, с четвертым – 12.

Средний прогнозируемый балл, который могут набрать испытуемые с разным уровнем подготовки, рассчитывался сложением баллов, набранных за счет имеющихся знаний и за счет угадывания ответов на оставшиеся вопросы с учетом вероятности правильного выбора одного из неисключенных вариантов ответа. Начиная с третьего уровня подготовки средний прогнозируемый балл превышает значение $T_{зач\,min}$, что говорит о том, что при отсутствии условий (3.17а) и (3.17б) вероятность ошибки второго

рода оказывается существенной. Влияние дополнительных условий на вероятность отказа ТОУД наглядно демонстрируется сравнением диаграмм на рис. 3.2.



а) б)

Рис. 3.2. Вероятность прохождения теста: а) – по критерию (3.14); б) – с учетом дополнительных условий (3.17а) и (3.17б)

Использование дополнительных условий (3.17а) и (3.17б) не только обеспечивает повышение надежности ТОУД за счет снижения вероятности ошибки второго рода, но и позволяет контролировать временную деградацию теста, например на основе сравнения процента результатов тестирования, удовлетворяющих критерию (3.14), и результатов, удовлетворяющих критерию (3.14б). Очевидно, увеличение процента испытуемых, преодолевающих критерий (3.14), не сопровождающееся таким же увеличением доли результатов, удовлетворяющим условиям (3.14), (3.17а) и (3.17б) одновременно, свидетельствует об аномально высоком проценте угадывания ответов на сложные вопросы, т. е. о необходимости реконструкции ТОУД. К аналогичным выводам можно прийти и на основе анализа соотношений правильных и неправильных ответов на вопросы различного уровня сложности. Для каждого теста нормальным является результат:

$$\frac{N_{легк}^{отв}}{N_{легк}^{вопр}} > \frac{N_{ср}^{отв}}{N_{ср}^{вопр}} > \frac{N_{тр}^{отв}}{N_{тр}^{вопр}}, \qquad (3.19)$$

а несоблюдение данного неравенства следует расценивать как проявление ошибки второго рода.

Повышение дифференцирующей способности теста, сформированного из ЕТЗ с выбором единственно верного ответа, за счет разделения вопросов на две и более категории сложности с применением критерия (3.14) в совокупности с дополнительными условиями (3.17а) и (3.17б) не только обеспечивает высокую достоверность оценок КТОЗУО, но и опровергает утверждение о снижении надежности и валидности теста при увеличении его дифференцирующей способности.

Использование дополнительных условий вида (3.17а), (3.17б), (3.19) позволяет диагностировать отказ КТОЗУО и поддерживать отказоустойчивое стояние теста путем своевременного изменения значений коэффициентов в выражениях (3.14), (3.17а), (3.17б).

3.2. Математическая модель и результаты оценки надежности теста из заданий с множественным выбором верных ответов

Широкое внедрение в образование компетентностного подхода к формированию и реализации образовательных программ послужило импульсом развития методик и программных средств оценивания учебных достижений обучающихся на основе тестов. Одной из наиболее сложных в дидактическом и техническом аспекте задач, решаемых при реализации образовательных программ на основе компетентностного подхода к обучению, является создание фонда оценочных средств для определения уровня сформированности компетенций [66, 67]. Оценка сформированности компетенций, предусмотренных ФГОС и основной профессиональной образовательной программой (ОПОП), наиболее актуальна для всех заинтересованных в результатах обучения лиц (обучающихся, специалистов образовательных организаций, работодателей) на завершающем этапе подготовки специалистов, в идеальном случае – накануне или в ходе Государственной итоговой аттестации (ГИА). Однако требования к содержанию программ ГИА и формам ее проведения обеспечивают проверку освоения обучающимися лишь наиболее важных, ключевых компетенций, предусмотренных ОПОП. Таким образом, одной из

основных проблем контроля качества обучения при реализации ФГОС является необходимость выполнения комплекса противоречивых требований к оценке сформированности следующих компетенций:

– объективность, основывающаяся на корректности (валидности) методик и содержания контроля;

– достоверность, необходимым условием обеспечения которой является комплексный характер контроля;

– актуальность, предполагающая проведение оценочных процедур в максимально сжатые сроки.

Не умаляя значимости развивающихся в последнее время интерактивных форм контроля знаний и умений обучающихся, основанных на выполнении практико-ориентированных задач, таких как учения, деловые игры и др. [68–70], следует отметить, что КТОЗУО в наибольшей мере удовлетворяют совокупности перечисленных выше требований [65, 71]. Их очевидными преимуществами перед другими формами контроля являются принципиальная возможность охвата всего содержания обучения, автоматизация процедур первичного анализа результатов, а также, особенно с применением сетевых технологий, оперативность проведения оценочных мероприятий с максимальным участием обучающихся.

Закрепление названных преимуществ требует совершенствования методик и алгоритмов компьютерного тестирования, направленного на повышение достоверности и объективности результатов контроля. Очевидным решением, обеспечивающим высокий уровень объективности и достоверности КТОЗУО, является увеличение объема тестовой выборки и базы заданий в целом путем дифференциации сложности заданий и одновременно достижением максимального тематического охвата содержания контроля.

Поскольку каждая компетенция формируется несколькими дисциплинами, будем полагать их среднее количество равным D, а среднее количество дидактических единиц учебной дисциплины, составляющих содержание формируемой компетенции, – E. При четырех уровнях оценки освоения каждой дидактической единицы количество вопросов и заданий с выбором единственно верного ответа в индивидуальной выборке КТОЗУО:

$$N_i = 4D \cdot E, \qquad (3.20)$$

а количество вопросов и заданий в базе заданий, обеспечивающее достаточную степень уникальности индивидуальной выборки при случайном выборе вопросов для каждой дидактической единицы:

$$N_Б = M \cdot N_i = 4M \cdot K \cdot D \cdot E, \quad (3.21)$$

где К – количество компетенций, предусмотренных ОПОП, М≥3.

Оценки, полученные по формуле (3.21) для различных ОПОП, дают значения $N_Б$ = 20000…50000, что свидетельствует о высокой трудоемкости разработки КТОЗУО из вопросов с выбором единственно верного ответа. Большие значения $N_Б$ повышают вероятность ошибок первого рода [61, 72], время устранения которых на этапе отладки теста и в начале его эксплуатации находится в прямой зависимости от $N_Б$.

Одним из направлений поиска решений, обеспечивающих уменьшение $N_Б$ при сохранении требуемых показателей надежности КТОЗУО, является усложнение структуры теста, включая:

– усложнение структуры ЕТЗ;

– усложнение алгоритма оценивания ЕТЗ;

– включение в базу заданий и индивидуальные тестовые выборки ЕТЗ с различной структурой и сложностью;

– усложнение алгоритма оценивания КТОЗУО.

Выше рассмотрен алгоритм оценивания КТОЗУО, составленного из вопросов с выбором единственно верного ответа, учитывающего уровень сложности вопроса и содержащего дополнительные условия, снижающие вероятность получения испытуемым положительной оценки за счет угадывания правильных ответов [65].

Для сравнения с приведенными в разделе 3.1 данной монографии результатами моделирования рассмотрим тест, составленный из ЕТЗ с множественным выбором верных ответов [9, 63, 73, 74]. Структура ЕТЗ в данном случае наиболее близка к заданию с выбором единственно верного ответа, однако возможные способы оценивания такого выбора намного разнообразнее.

Рассмотрим ЕТЗ в виде вопроса с N вариантами ответа, из них K правильных и, соответственно, $N – K$ неправильных. Будем полагать, что испытуемый знает m верных вариантов ответа ($0 \le m \le K$ -1), а оставшиеся $n = K – m$ ответов угадывает. Вероятность угадать ровно k ($0 \le k \le n$) верных ответов [46]:

$$P_{\text{угад}} = \frac{C_K^k \cdot C_{N-K}^{n-k}}{C_N^n}, \quad (3.22)$$

где величины вида C_N^n рассчитываются:

$$C_N^n = \frac{N!}{n!(N-n)!}. \quad (3.23)$$

Например, вопрос имеет пять вариантов ответа, из них три правильных и два неправильных. Полагаем, что испытуемый знает один правильный ответ. Тогда вероятность угадать еще два правильных ответа из четырех оставшихся:

$$P_{\text{угад}} = \frac{C_2^2 \cdot C_2^0}{C_4^2} = 0{,}166(6).$$

В отличие от ЕТЗ с выбором единственно верного ответа в данном случае можно рассмотреть несколько способов оценивания ответа с поощрением частично верных ответов и штрафов за выбор неверных вариантов ответа [63]. Нам представляется корректной следующая систематизация способов оценивания ЕТЗ:

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.

Так же, как и в [65], представим модель испытуемого шестью уровнями подготовки, которым соответствует знание 0, 20, 40, 60, 70 и 80 процентов правильных ответов соответственно. Однако в отличие от [65] детализируем модели дополнительными уровнями, характеризующимися возможностью, полагаясь на свои знания, исключить из случайного выбора некоторое количество верных (в) или неверных (нв) вариантов ответа (табл. 3.3).

В таблице 3.3 представлены результаты моделирования решения теста, состоящего из 20 вопросов с несколькими вариан-

тами верных ответов (10 вопросов с $N = 4$, $K = 2$; 5 вопросов с $N = 5$, $K = 2$; 5 вопросов с $N = 5$, $K = 3$), ЕТЗ считалось выполненным и один балл начислялся за K верных ответов в K выбранных вариантах (1-й способ оценивания ЕТЗ).

Таблица 3.3

Результаты компьютерного моделирования решения теста А

Номер модели	Уровень подготовки	Категория вопроса	% известных вопросов	Количество известных вопросов	Количество исключаемых вариантов ответа	Вероятность угадывания	Гарантированные баллы	Недостающие баллы	Вероятность получения положительной оценки	Количество набранных баллов за счет угадывания	Средний прогнозируемый балл
1	1	4(2)	0	0	0	0,1667	0	16	$1{,}94\cdot10^{-11}$	1,6667	2,667
		5(2)		0		0,1				0,5	
		5(3)		0		0,1				0,5	
2		4(2)		0	1 в	0,3333			$8{,}60\cdot10^{-7}$	3,3333	5,417
		5(2)		0		0,25				1,25	
		5(3)		0		0,1667				0,8333	
3		4(2)		0	1 н/в	0,3333			$8{,}60\cdot10^{-7}$	3,3333	5,417
		5(2)		0		0,1667				0,8333	
		5(3)		0		0,25				1,25	
4	2	4(2)	20	2	1 в	0,3333	4	12	$7{,}06\cdot10^{-5}$	2,6667	8,333
		5(2)		1		0,25				1	
		5(3)		1		0,1667				0,6667	
5		4(2)		2	1 н/в	0,3333			$7{,}06\cdot10^{-5}$	1,3333	8,333
		5(2)		1		0,1667				0,3333	
		5(3)		1		0,25				0,1667	
6		4(2)		2	1 в и 1 н/в	0,5			0,0065	4	10,67
		5(2)		1		0,3333				1,3333	
		5(3)		1		0,3333				1,3333	
7	3	4(2)	40	4	1 в	0,3333	8	8	0,0043	2	11,25
		5(2)		2		0,25				0,75	
		5(3)		2		0,1667				0,5	
8		4(2)		4	1 н/в	0,3333	8	8	0,0043	2	11,25
		5(2)		2		0,1667				0,5	
		5(3)		2		0,25				0,75	
9		4(2)		4	1 в и 1 н/в	0,5	8	8	0,0695	3	13
		5(2)		2		0,3333				1	
		5(3)		2		0,3333				1	
10	4	4(2)	60	6	1 в	0,3333	12	4	0,1416	1,3333	14,17
		5(2)		3		0,25				0,5	
		5(3)		3		0,1667				0,3333	

Окончание табл. 3.3

11		4(2)		6	1 н/в	0,3333	12	4	0,1416	1,3333	14,17
		5(2)		3		0,1667				0,3333	
		5(3)		3		0,25				0,5	
12		4(2)		6	1 в и 1 н/в	0,5	12	4	0,4444	2	15,33
		5(2)		3		0,3333				0,6667	
		5(3)		3		0,3333				0,6667	
13	5	4(2)	70	7	1 в и 1 н/в	0,5	13	3	0,5925	1,5	15,83
		5(2)		3		0,3333				0,6667	
		5(3)		3		0,3333				0,3333	
14	6	4(2)	80	8	1 в и 1 н/в	0,5	16	0	1	1	17,67
		5(2)		4		0,3333				0,3333	
		5(3)		4		0,3333				0,3333	

Для оценки влияния структуры теста на положительный результат тестирования в табл. 3.4 приведены результаты компьютерного моделирования решения теста Б, также состоящего из 20 вопросов: 6 вопросов с $N = 4, K = 2$; 7 вопросов с $N = 5, K = 2$; 7 вопросов с $N = 5, K = 3$.

Таблица 3.4

Результаты компьютерного моделирования решения теста Б

Номер модели	Категория вопроса	Количество известных вопросов	Гарантированные баллы	Недостающие баллы	Вероятность получения положительной оценки	Количество набранных баллов за счет угадывания	Средний прогнозируемый балл
4	4(2)	1	3	13	$7{,}207 \cdot 10^{-6}$	1,6667	7,167
	5(2)	1				1,5	
	5(3)	1				1	
5	4(2)	1	3	13	$7{,}207 \cdot 10^{-6}$	1,6667	7,167
	5(2)	1				1	
	5(3)	1				1,5	
6	4(2)	1	3	13	0,0014	2,5	9,5
	5(2)	1				2	
	5(3)	1				2	
7	4(2)	2	8	8	0,0024	1,333	11
	5(2)	3				1	
	5(3)	3				0,6667	
8	4(2)	2	8	8	0,0024	1,333	11
	5(2)	3				0,6667	
	5(3)	3				1	

Окончание табл. 3.4

9	4(2)	2	8	8	0,0462	2	12,67
	5(2)	3				1,333	
	5(3)	3				1,333	
10	4(2)	6	12	4	0,0465	1	13,25
	5(2)	3				0,75	
	5(3)	3				0,5	
11	4(2)	6	12	4	0,0465	1	13,25
	5(2)	3				0,5	
	5(3)	3				0,75	
12	4(2)	6	12	4	0,2407	1,5	14,5
	5(2)	3				1	
	5(3)	3				1	
13	4(2)	4	13	3	0,5391	1	15,67
	5(2)	4				1	
	5(3)	5				0,6667	
14		16	16	0	1	1,5	17,5

В таблице 3.4 не приведены результаты для моделей обучающихся № 1–3 ввиду пренебрежимо малой вероятности прохождения теста при нулевом проценте знаний. Для моделей обучающихся № 7–11 с учетом округления до целочисленных значений количества известных вопросов сопоставление условий тестирования наиболее корректно. Как видно из сопоставления данных, приведенных в табл. 3.5 и 3.4, изменение структуры теста в наименьшей мере влияет на результат тестирования в сравнении с количеством вопросов теста и значением порога положительной оценки.

Для модели обучающегося № 12 при решении теста А получен средний прогнозируемый балл с учетом возможного угадывания 15,33, «пограничный» с удовлетворительным результатом, что дало нам основание провести дополнительное исследование на предмет влияния способов оценивания ЕТЗ на итоговый балл.

Вычитание $1/(N - K)$ балла (K – количество верных ответов из N предлагаемых) за каждый выбранный неверный ответ (3-й способ оценивания ЕТЗ) привело к снижению среднего прогнозируемого балла более чем на 1, до 14,17, а вероятности прохождения теста – на целый порядок, до 0,0579.

Прибавление $1/K$ балла за выбор каждого верного варианта ответа в дополнение к штрафу за выбор неверных ответов (4-й способ оценивания ЕТЗ) обеспечило превышение средним про-

гнозируемым баллом (16,33) уровня удовлетворительной отметки, вероятность прохождения теста возросла до 0,6.

Сравнение с результатами компьютерного моделирования, приведенными в подразделе 3.1, свидетельствует о том, что структура ЕТЗ с несколькими верными вариантами ответа позволяет повысить компактность КТОЗУО, т. е. потенциально обеспечивает требуемый уровень достоверности оценки знаний и умений испытуемого при существенно меньшем в сравнении с ЕТЗ с выбором единственно верного ответа количеством заданий.

Одним из направлений совершенствования рассмотренной структуры КТОЗУО и основанной на нем методики проверки знаний и умений, на наш взгляд, может быть контроль надежности теста, т. е. выявление с большой степенью достоверности и предотвращение заучивания правильных ответов, например получаемых от испытуемых, ранее прошедших тестирование, без их понимания. Для этого, по аналогии с [65], КТОЗУО должен содержать некоторое количество относительно несложных вопросов и заданий, а критерий положительной оценки должен быть дополнен условием превышения количественного порога ответов на такие вопросы, причем процент ответов на несложные вопросы (задания) должен существенно превышать процент ответа на все вопросы и задания КТОЗУО. Невыполнение этого условия не только не позволяет оценить решение теста положительно даже при превышении общего порогового значения правильно выполненных заданий, но и свидетельствует о «взломе» теста и необходимости его замены. Включение в тест, составленный из ЕТЗ с выбором единственно верного варианта ответа, заведомо несложных вопросов и заданий при сохранении требуемого уровня достоверности оценки знаний и умений обучающихся, неизбежно влечет за собой увеличение объема теста или повышение порога положительной оценки, что потенциально увеличивает вероятность ошибок первого рода [61, 62, 72]. Однако структура ЕТЗ с множественным выбором верных ответов обеспечивает контроль надежности КТОЗУО при сохранении неизменным объема теста. Для этого один из вариантов ответа на вопрос (задание) следует сделать относительно простым для испытуемого и контролировать количество выборов таких ответов или их процент в общем

количестве выбранных (или только выбранных правильных) ответов.

Примеры таких вопросов.

1. Современниками А.С. Пушкина были:

а) А.А. Бестужев-Марлинский;

б) А.М. Иванов-Крамской;

в) Н.И. Миклухо-Маклай;

г) С.И. Муравьев-Апостол;

д) С.Н. Сергеев-Ценский.

2. Сопротивление металлического проводника возрастает при увеличении:

а) длины свободного пробега электрона в его объеме;

б) порядкового номера металла в периодической системе химических элементов;

в) частоты тока;

г) температуры;

д) концентрации неметаллических примесей.

Из двух правильных ответов – «а» и «г» – на первый вопрос ответ «г» относительно прост, так как многие знают, что Пушкин – современник участников Декабрьского восстания 1825 г., а Муравьев-Апостол – один из казненных лидеров декабристов.

Из трех правильных ответов – «в», «г» и «д» – на второй вопрос ответ «г» наиболее известен и наиболее часто воспроизводим при изучении электричества.

3.3. Моделирование надежности теста из заданий на установление парных соответствий

Расширенная реализация в дистанционном формате образовательных программ, обусловленная в числе прочего мерами профилактики распространения новых видов инфекций, активизировала совершенствование методик дистанционного обучения, внедрение в учебный процесс специализированных программных продуктов, в том числе предназначенных для проведения занятий в офлайн-режиме и предполагающих оценивание знаний и умений обучающихся. Сложности организации и проведения таких занятий во многом обусловлены противоречивостью дидактиче-

ских 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;

– необходимостью совмещения обучающего и контролирующего компонентов при ограничении времени доступа обучающегося к образовательному ресурсу;

– снижением или ограничением трудоемкости разработки и переработки офлайн-занятий наряду с необходимостью унификации и оптимизации их структуры, обеспечения возможности многократного использования отдельных занятий для различных образовательных программ;

– соблюдением единства требований и критериев оценивания знаний обучающихся по различным учебным дисциплинам, обеспечением объективности оценок.

Типичная структура обучающего ресурса, например темы в электронном учебном пособии или обособленного обучающе-контрольного материала, включает в себя: план изучаемого материала (перечень ДЕСО); содержательно-обучающий контент (текстовые, видео-, аудио- и графические информационные материалы, в том числе примеры решения задач, анализ типовых ошибок и т. п.); вопросы и задания для самоконтроля; оценочные (контрольно-измерительные) материалы (задачи, тесты и др.). Многократное использование такого рода ресурсов обостряет проблему надежности тестов для оценивания результатов обучения за счет увеличения вероятности ошибок второго рода [61, 72, 75]. Как показывают качественные оценки [9, 63, 73, 76] и результаты моделирования надежности тестов [63, 65, 77, 78], оптимизация объема теста основывается на компромиссе между обеспечением его высокой надежности, требующим увеличения количества тестовых заданий и вариативности их выборок за счет создания обширных баз заданий и их регулярного пополнения или обновления, и ограничением времени тестирования. Наряду с минимизацией ошибок второго рода, увеличение объемов теста, в том числе за счет дифференцирования сложности заданий, нередко

преследует цель детализации анализа результатов тестирования: дифференциации по степени усвоения материала отдельных тем или ДЕСО, выявления динамики забывания учебного материала, сравнения результатов учебных достижений различных групп обучающихся и др. [79–82].

Методики автоматизированного адаптивного контроля знаний, в особенности основанные на модели Раша [63, 83], позволяющие детализировать оценки учебных достижений обучающихся и реализованные, в частности, в тестах с уточняющими вопросами, цепочками вопросов, с варьированием сложности заданий содержанием и подбором вариантов ответа и др., предполагают как использование обширных баз заданий, так и относительно больших временных интервалов тестирования [83–87].

Решения, обеспечивающие компактность компьютерных тестов для оценки учебных достижений обучающихся и, соответственно, оперативность тестирования при приемлемой достоверности оценок, на наш взгляд, должны основываться на повышении надежности (отказоустойчивости, вероятности безотказной работы) ЕТЗ, прежде всего за счет разработки или выбора структуры ЕТЗ, максимально устойчивой к деградации (механического запоминания правильных ответов при многократном использовании как индивидуально, так и группой тестируемых) и угадыванию правильного ответа. Так, тесты, составленные из ЕТЗ с выбором единственно верного ответа, оказываются наиболее подверженными деградации и, как следствие, ошибкам второго рода при оценивании знаний обучающихся. Сравнение результатов моделирования надежности тестов, составленных из ЕТЗ с выбором единственно верного ответа и ЕТЗ с множественным выбором [65, 77], приведенных в подразделах 3.2 и 3.3 настоящей монографии, позволяет сделать вывод о том, что усложнение структуры ЕТЗ позволяет в несколько раз повысить компактность теста при сохранении приемлемых малых значений вероятности угадывания правильных ответов, в основном определяющей вероятность ошибок второго рода как основного вида отказов тестов, т. е. надежность теста в целом.

Разнообразие видов ЕТЗ в тесте, включая ЕТЗ с усложненной структурой и ЕТЗ открытого типа [9, 63, 79], требует от испытуемых дополнительной мобилизации, умения быстро пере-

страивать мыслительные процессы и способствует повышению объективности оценок знаний обучающихся, одновременно значительно расширяя творческий инструментарий преподавателей – разработчиков КТОЗУО.

Следует признать, что уменьшение количества заданий теста за счет усложнения структуры ЕТЗ не приводит к пропорциональному уменьшению времени решения теста, так как ЕТЗ со сложной структурой требуют большего времени на решение. В связи с этим представляет интерес сопоставление по единой методике показателей надежности тестов с различной структурой ЕТЗ для оптимизации структуры тестов, включающей распределение заданий (вопросов) по видам ЕТЗ, последовательность и временные лимиты тестирования.

Для обеспечения корректности сопоставления результатов моделирование надежности КТОЗУО, сформированных из одинаковых ЕТЗ различных типов, основывается на следующих положениях:

1) КТОЗУО рассматривается в качестве отбраковочного (для двухбалльной шкалы оценок – «зачет-незачет») или сортировочного (для многобалльной шкалы оценок) автомата, что позволяет применять к исследованиям КТОЗУО методологический и математический аппарат теории надежности [47, 61];

2) разная степень обученности испытуемых моделируется несколькими уровнями подготовки, характеризующимися различными процентами известных испытуемым ответов от их общего количества в зависимости от структуры ЕТЗ;

3) применяются различные алгоритмы пересчета верных и неверных ответов, в разной степени учитывающие структуру ЕТЗ назначением баллов за частично верные ответы;

4) для сравнения величин ошибок второго рода при применении тестов на основе ЕТЗ одного вида рассматривается двухбалльная шкала оценок («зачет-незачет»);

5) уровень удовлетворительной оценки устанавливается одинаковым для всех анализируемых структур тестов и видов ЕТЗ и составляет 80 % от максимально возможного количества набранных баллов [65, 77].

Рассмотрим относительно редко используемое при создании тестов, но реализуемое во многих системах проверки знаний

(например, в популярной виртуальной обучающей среде Moodle) ЕТЗ на установление парных соответствий (ЕТЗУС) между множествами понятий, такими как «green, blue, gray, brown – серый, голубой, коричневый, зеленый», «Обь, Лена, Печора, Северная Двина – Белое, Баренцево, Карское, Лаптевых», «индуктивность, проводимость, емкость, сопротивление – Сименс, Фарада, Ом, Генри» и т. п.

При расчете вероятности получения положительной оценки («зачет») будем руководствоваться выражениями (3.6) – (3.9).

Вероятность угадывания полного правильного ответа «вслепую» для ЕТЗУС в общем случае [46]:

$$P_{\text{угад с0}} = \frac{1}{n_1!} \cdot \frac{1}{n_2!}, \tag{3.24}$$

где n_1 и n_2 – количество элементов в первом и втором множестве, а при их равенстве:

$$P_{\text{угад с0}} = (n_1!)^{-2} = (n_\text{с}!)^{-2}, \tag{3.25}$$

с – количество соответствий в ЕТЗУС.

Если испытуемому известно некоторое количество верных соответствий $n_\text{и}$, вероятность угадывания:

$$P_{\text{угад си}} = [(n_\text{с} - n_\text{и})!]^{-2}. \tag{3.26}$$

Будем полагать, что все соответствия взаимоисключающие, т. е. одному элементу первого множества соответствует только один элемент второго множества и наоборот, исключая задания с множествами типа: «Обь, Лена, Печора, Енисей – Белое, Баренцево, Карское, Лаптевых» или «green, blue, gray, brown – серый, голубой, черный, зеленый».

Тогда вероятность угадывания полного правильного ответа «вслепую»:

$$P_{\text{угад с0}} = (n_\text{с}!)^{-1}, \tag{3.25а}$$

а при известном испытуемому количестве $n_\text{и}$ связанных пар:

$$P_{\text{угад си}} = [(n_\text{с} - n_\text{и})!]^{-1}. \tag{3.26а}$$

По аналогии с условиями моделирования, представленными в разделе 3.2 [77], рассмотрим тест из двадцати ЕТЗ примерно одинаковой сложности, по четыре парных соответствия в каждом задании. Тогда $P_{\text{угад с0}} = 0{,}0416667$; $P_{\text{угад с1}} = 0{,}1666667$; $P_{\text{угад с2}} = 0{,}5$.

Представим модель испытуемого шестью уровнями подготовки (табл. 3.5).

Таблица 3.5

Характеристики модели испытуемого

Уровень подготовки	Процент ЕТЗ с известными соответствиями			Гарантированно набираемые баллы	Недостающие баллы
	с одним	с двумя	с тремя или четырьмя		
0	0	0	0	0	16
1	50	25	0	0	16
2	20	40	20	4	12
3	20	30	40	8	8
4	20	30	50	10	6
5	10	30	60	12	4

Знание четырех соответствий ЕТЗ означает его верное решение и гарантированное получение балла. Знание трех соответствий ЕТЗ из четырех автоматически приводит к верному решению.

В таблице 3.6 приведены результаты расчетов вероятности получения удовлетворительной оценки и средних набираемых испытуемыми баллов.

Таблица 3.6

Результаты компьютерного моделирования решения теста из двадцати ЕТЗУС

Уровень подготовки	Вероятность прохождения теста	Количество баллов за счет угадывания	Прогнозируемый средний набранный балл
0	$3{,}4075 \cdot 10^{-19}$	0,833	0,833
1	$1{,}2119 \cdot 10^{-9}$	2,083	2,083
2	$2{,}83 \cdot 10^{-5}$	4,375	4,375
3	0,0044	4,833	8,833
4	0,1003	3,75	11,75
5	0,4457	3,333	15,333

Полученные данные свидетельствуют о том, что даже относительно высокий уровень подготовки (3/4 от минимально доста-

точного количества $M_{зач}$ априори правильно решенных ЕТЗУС и 3/8 от $M_{зач}$ наполовину известных решений ЕТЗУС) далеко не гарантирует положительной оценки, а при $M_{зн} \leq 0{,}6 \cdot M_{зач}$ вероятность ее получения становится пренебрежимо малой. Для 3–5 уровней подготовки испытуемых вероятность ошибки второго рода меньше, чем дает тест, составленный из ЕТЗ с множественным выбором верных ответов, и немногим больше, чем обеспечивает тот же тест с усложненным алгоритмом оценивания, предусматривающим штрафы за выбор неверных вариантов ответов [77].

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

$$I_0 = \sum_{i=2}^{n_с} i \leq n_с!$$

Однако уже при $n_с = 3$ ЕТЗУС в алгоритмах тестирования с подсказками обеспечивает меньше вероятность ошибки второго рода, чем ЕТЗ с выбором одного верного варианта ответа из четырех, что обеспечивает КТОЗУО на основе ЕТЗУС дополнительный дидактический потенциал. Условием получения удовлетворительной оценки при решении теста с подсказками может быть:

$$\frac{П}{П_{зач}} = \frac{П_в + П_н}{П_{зач}} \leq 1, \qquad (3.27)$$

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

3.4. Моделирование надежности теста из заданий на установление правильной последовательности

Единичные тестовые задания на установление правильной последовательности понятий, терминов, символов и др. (ЕТЗУП) применяются разработчиками КТОЗУО сравнительно редко, что объясняется относительной сложностью формулирования ЕТЗУП и явной предпочтительностью ЕТЗ других типов для проверки знаний содержания отдельных учебных дисциплин и ДЕСО. С другой стороны, для оценивания учебных достижений по дисциплинам, использующим символьные выражения, устойчивые лингвистические конструкции, сформировавшийся однозначный понятийный аппарат, наработан большой методический и эмпирический опыт применения такого специфического вида ЕТЗ. В целях простоты реализации проверочных алгоритмов элементы таких последовательностей нумеруются по исходному порядку, и ответ на ЕТЗ представляет собой последовательность цифр с верным порядком следования элементов сгенерированной случайным образом последовательности или ее части.

Пример 1: «сила притяжения зарядов прямо пропорциональна; обратно пропорциональна; квадрат (вторая степень); диэлектрическая проницаемость среды взаимодействия; величина первого заряда; величина второго заряда».

Пример 2: «is; task; it; difficult; not».

Пример 3: «укажите последовательность операций в логическом выражении: $a \to b \wedge c \leftrightarrow d \vee e$; конъюнкция, дизъюнкция, импликация, эквиваленция».

Вероятность угадывания правильной последовательности «вслепую», как и в случае ЕТЗУС с взаимоисключающими соответствиями:

$$P_{\text{угад п0}} = (n_{\text{п}}!)^{-1}\,, \qquad (3.25б)$$

где $n_{\text{п}}$ – количество элементов последовательности. В ряде последовательностей во избежание нескольких формально правильных вариантов ответа одна (обычно первая) или две позиции предопределяются логикой построения последовательности, являясь ее отправными или связующими элементами. Таковой в примере 1 является первая позиция выражения «сила притяжения

зарядов прямо пропорциональна». Если разделить это выражение на две части: «сила притяжения зарядов» и «прямо пропорциональна», то в примере 1 может быть несколько формально правильных (неошибочных) вариантов ответа. Поэтому в общем случае вероятность правильного решения ЕТЗУП:

$$P_{\text{угад пи}} = \left[(n_{\text{п}} - n_{\text{ф}} - n_{\text{и}})!\right]^{-1}, \qquad (3.26\text{б})$$

где $n_{\text{ф}}$ и $n_{\text{и}}$ – количество фиксированных (определяемых структурой и логикой построения последовательности) и заранее известных испытуемому позиций элементов последовательности.

Таким образом, отличия в моделировании надежности КТОЗУО, составленных из ЕТЗУС или ЕТЗУП, заключаются в учете априори известных испытуемым сведениям, т. е. в различиях моделей испытуемых. Будем полагать, что КТОЗУО составлен из двадцати ЕТЗУП двух типов: тип I – последовательность из четырех элементов; тип II – последовательность из пяти элементов с двумя фиксированными позициями, по десять ЕТЗ каждого типа.

С учетом особенностей выполнения ЕТЗУП модель испытуемого представлена семью уровнями подготовки:

– нулевой: испытуемый решает тест «вслепую» без каких-либо знаний;

– первый: испытуемый знает по 10 % ответов на ЕТЗ обоих типов, а также правильную позицию одного элемента в 20 % ЕТЗ обоих типов;

– второй: испытуемый знает по 20 % ответов на ЕТЗ обоих типов, а также правильные позиции элементов последовательностей: одного – в 20 % заданий типа I и 30 % заданий типа II, двух – в 20 % заданий типа I;

– третий: испытуемый знает по 30 % ответов на ЕТЗ обоих типов, правильные позиции элементов последовательностей: одного – в 30 % заданий типа I и 50 % заданий типа II, двух – в 30 % заданий типа I;

– четвертый: знание испытуемым по 40 % ответов на ЕТЗ обоих типов, а также правильных позиций элементов последовательностей: одного – в 40 % заданий типа I и 50 % заданий типа II, двух – в 20 % заданий типа I;

– пятый: знание испытуемым по 50 % ответов на ЕТЗ обоих типов и правильных позиций элементов последовательностей:

одного – в 20 % заданий типа I и 50 % заданий типа II, двух – в 30 % заданий типа I;

– шестой: знание испытуемым по 60 % ответов на ЕТЗ обоих типов, а в оставшихся 40 % ЕТЗ знание правильных позиций элементов последовательностей: одного – в заданиях типа II и двух – в заданиях типа I.

В таблице 3.7 приведены результаты расчетов среднего балла, набираемого за счет угадывания, общего среднего балла и вероятности получения удовлетворительной оценки для данных моделей теста и испытуемых.

Таблица 3.7

Результаты компьютерного моделирования решения теста из двадцати ЕТЗУП

Уровень подготовки	Гарантированно набираемые баллы	Недостающие баллы	Вероятность получения удовлетворительной оценки	Количество баллов за счет угадывания	Прогнозируемый набранный балл
0	0	16	$4{,}3063 \cdot 10^{-14}$	2,0833	2,0833
1	2	14	$2{,}3837 \cdot 10^{-10}$	2,7917	4,7917
2	4	12	$1{,}0615 \cdot 10^{-6}$	3,8333	7,8333
3	6	10	0,0016	4,875	10,875
4	8	8	0,0118	4,333	12,333
5	10	6	0,2190	4,333	14,333
6	12	4	0,6367	4,0	16

Сравнение данных, приведенных в табл. 3.6 и 3.7, показывает, что наличие в тесте ЕТЗ типа II увеличивает вероятность ошибок второго рода за счет угадывания правильных ответов. Повышение надежности КТОЗУО, составленных из ЕТЗУП, может быть достигнуто за счет увеличения количества позиций в ЕТЗУП и исключения последовательностей с фиксированными позициями элементов. Однако, как показывает наш опыт применения таких

тестов, увеличение позиций последовательности до шести и более приводит к существенному, в два раза и более по сравнению с рассматриваемыми моделями ЕТЗУП, увеличению времени решения заданий, в том числе за счет необходимости тщательной предварительной проверки испытуемыми ответа перед его вводом. К тому же усложнение ЕТЗУП за счет увеличения числа комбинаций элементов приводит к наибольшему среди рассмотренных нами в данном разделе монографии ЕТЗ увеличению вероятности ошибок первого рода, порожденных как ошибками при вводе ответа, так и ошибками разработчиков КТОЗУО.

Сопоставление результатов моделирования надежности тестов, составленных из ЕТЗ с усложненной структурой и представленных в разделах 3.2–3.4, позволяет сделать следующие выводы:

– вероятность угадывания верного ответа на ЕТЗ с множественным выбором верных ответов, ЕТЗУС и ЕТЗУП существенно меньше, чем на ЕТЗ с выбором единственно верного ответа, что позволяет формировать из названных ЕТЗ с усложненной структурой относительно компактные и надежные КТОЗУО;

– вероятности угадывания верного ответа для всех трех рассмотренных нами типов ЕТЗ с усложненной структурой близки, что позволяет комбинировать ЕТЗ данных видов в КТОЗУО, заменяя один другим, без необходимости повторных оценок надежности КТОЗУО;

– ЕТЗУП и, в особенности ЕТЗУС, могут быть использованы для контроля надежности (обнаружения «взлома») КТОЗУО на идейной основе метода Раша. Для этого в тесте может содержаться некоторое количество сравнительно несложных ЕТЗ, компенсируемое значением порога положительной оценки, а в критерии положительной оценки – условие превышения значения порога правильных ответов на несложные ЕТЗ. Дополнительное условие критерия положительной оценки может основываться и на том, что соотношение верных и неверных ответов на несложные ЕТЗ должно значительно превышать аналогичное соотношение для оставшихся (более сложных) ЕТЗ КТОЗУО или всего КТОЗУО в целом:

$$\frac{M_{пр\,л}}{N_л - M_{пр\,л}} \geq k_1 \cdot \frac{M_{пр}}{N_л - M_{пр}}; \qquad (3.28)$$

$$\frac{M_{пр\,л}}{N_л - M_{пр\,л}} \geq k_2 \cdot \frac{M_{пр} - M_{пр\,л}}{N - N_л - (M_{пр} - M_{пр\,л})}. \quad (3.29)$$

Здесь $N_л$, $M_{пр\,л}$ – количество легких ЕТЗ в КТОЗУО и ответов на них. Определяющие различие в сложности ЕТЗ коэффициенты k_1 и k_2 в выражениях (10) и (11) могут быть определены эмпирическим или экспертным путем;

– ЕТЗ с усложненной структурой допускают частично верные ответы: установление части верных соответствий в ЕТЗУС, правильное указание отдельных позиций элементов последовательности в ЕТЗУП, выбор одного из правильных вариантов ответа в ЕТЗ с множественным выбором, что дает возможность, во-первых, строить на их основе сложные оценочные алгоритмы и шкалы, во-вторых, получать детализированные результаты тестирования для анализа индивидуальных и групповых учебных достижений применительно к различным предметным областям, ДЕСО и компетенциям на основе сравнительно компактных по объему ЕТЗ и непродолжительных по времени исследований.

4. Некоторые вопросы повышения надежности и достоверности тестирования учебных достижений

4.1. Повышение дидактического потенциала тестирования учебных достижений

Разработчик теста во избежание превышения вероятностью ошибок второго рода некоторого определенного им критического значения, зависящего от порога положительной оценки (при двухбалльной шкале оценивания «зачет-незачет») или интервалов между порогами оценок (при многобалльной шкале оценивания), должен регулярно модифицировать тест, и частота таких модификаций определяется оценкой надежности теста [75]. В свою очередь, потенциальная устойчивость теста к ошибкам второго рода определяется его структурой: распределением заданий по уровням сложности, их взаимной корреляцией, обеспечивающей дополнительную верификацию оценок, а также типом и конструкцией ЕТЗ. Снижению вероятности ошибок второго рода способствует формирование теста из ЕТЗ, предполагающих множественный выбор верных ответов, установление соответствия между объектами и их признаками или группами признаков, установление правильной последовательности, например слов в определении или формулировке.

Пример ЕТЗУП – формулировка закона Ома для участка цепи:

1) сопротивление;
2) отношение;
3) равно;
4) ток;
5) напряжение.

Правильная комбинация цифр (вводимое число) – 43251, допустимый ответ – 13254. С учетом того, что «3» и «2» на второй и третьей позиции соответственно находятся постоянно, вероятность угадывания правильной комбинации – 1/6, за вычетом допустимого ответа – 1/5. Уменьшить вероятность угадывания в ЕТЗУП можно за счет увеличения количества позиций (разрядов числа) и уменьшения количества фиксированных позиций, определяемых структурой (правильным порядком слов) формулиров-

ки, однако далеко не каждое правило (формула, закон и т. п.) пригодно для представления заданием с установлением (восстановлением) правильной последовательности.

Применение ЕТЗ такого рода значительно увеличивает трудоемкость создания теста и требует относительно высокой методической квалификации разработчика, однако, по мнению критиков идеи проверки знаний и умений с помощью тестов, способствует лишь более достоверному выявлению знания ответов на отдельные вопросы, но не понимания изученного материала.

В работах [88, 92] нами предложена конструкция ЕТЗ, по форме ответа (вводимого числа) тождественная установлению правильной последовательности или градации ответов, однако, на наш взгляд, обеспечивающая проверку не только фрагментарных знаний, но и понимания изученного материала и обладающая несопоставимо большими дидактическими возможностями по сравнению с наиболее распространенными конструкциями ЕТЗ. В основе формирования последовательности – распределение предлагаемых вариантов ответа по степени близости к правильному, т. е. номер правильного ответа заносится в первый разряд, номер наиболее близкого к правильному ответу дистрактора – во второй и т. д.

Рассмотрим пример.

«Закон Ома для участка цепи:

1) $E = m/c^2$;

2) $Q = I^2 \cdot R \cdot t$;

3) $I = U/R$;

4) $E = A/t$;

5) $I = E/(R + r)$».

Первую позицию вводимого числа будет занимать цифра «3». Ближайший к нему по достоверности – ответ под цифрой «5» – закон Ома для цепи с источником ЭДС. На третьей позиции стоит расположить ответ под номером «2» – одну из записей закона Джоуля-Ленца, объясняя свой выбор, во-первых, тем, что запись закона правильная, а во-вторых, тем, что из оставшихся трех уравнений это единственное имеет отношение к электрическому току. Из оставшихся двух уравнений следует выбрать то, что записано без ошибки. Итоговый ответ на вопрос: 35214.

В отличие от ЕТЗ с установлением последовательности в данном наборе цифр отсутствуют фиксированные, определяемые правильным порядком слов в предложении позиции, что снижает вероятность угадывания правильного ответа до 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 всех вариантов ответов, приводимых в тесте в виде формул, графических объектов, предложений и их пересказ от испытуемого к испытуемому с нарастающим объемом представляется задачей, по трудоемкости сопоставимой с добросовестным изучением учебной дисциплины или ее раздела, темы, ДЕСО, содержание которых подвергается проверке.

По своей структуре и способу оформления решения задание в приведенном выше примере похоже на тестовое задание с градуированными ответами, рассмотренное в разделе 1.2. Однако принципы составления и решения такого задания коренным образом отличаются от задания с градуированными ответами, что позволяет выделить рассматриваемое ЕТЗ в отдельный вид – задание с упорядочиванием ответом. Его основным отличием от задания с градуированными ответами является наличие дистракторов – кроме одного, единственно верного, все остальные варианты ответа являются неверными. Решение задания с упорядочиванием ответов заключается в определении сначала верного ответа, что роднит этот вид ЕТЗ с заданием с выбором единственно верного ответа, а затем в расположении дистракторов в порядке убывания от истины. Конструирование ЕТЗ основано на подборе или формулировании разработчиком теста набора дистракторов, характеризующихся разной степенью принадлежности к соответствующим ДЕСО, предметной области, отрасли знаний.

Для того чтобы продемонстрировать дидактический потенциал ЕТЗ с упорядочиванием ответов (ЕТЗУО), рассмотрим следующий пример.

«Как вычислить площадь круга S, если известен его радиус r?

1) $S = 2\pi \cdot r$;

2) $S = \pi \cdot r^2/4$;

3) $S = 4\pi \cdot r^3/3$;
4) $S = 4\pi^2$;
5) $S = \pi \cdot r^2$».

После выбора верного ответа под номером «5» испытуемый решает не менее сложную задачу, так как она требует логических рассуждений. Из оставшихся вариантов лишь один, под номером «2», в правой части дает размерность площади, поэтому цифра «2» должна занять второй разряд вводимого в качестве ответа пятизначного числа. Из оставшихся трех неправильных ответов единственный, под номером «1», имеет отношение к кругу, так как позволяет вычислить его периметр (длину окружности). Наконец, из вариантов ответа под номерами «3» и «4» один вообще не содержит радиус, т. е. является абсурдным и должен быть соотнесен с последним разрядом вводимого числа. Итог: 52134. В качестве исключаемого в первую очередь путем логических рассуждений может применяться и вариант ответа вида: $S = \pi/r^2$, т. е. с обратной зависимостью площади от радиуса.

Рассмотрим следующий пример.

«Расположите фигуры в порядке убывания отношения их периметра к площади:

1) круг;
2) квадрат;
3) правильный шестиугольник;
4) равносторонний треугольник;
5) равнобедренный неравносторонний треугольник».

В процессе получения правильного ответа (13245) испытуемый должен выявить как минимум две закономерности: 1) чем меньше углов у равностороннего выпуклого многоугольника, тем меньше отношение его периметра к площади; 2) при условии равенства количества углов отношение периметра к площади у неравностороннего многоугольника меньше, чем у равностороннего. Получить правильный ответ на ЕТЗ и выявить две эти закономерности испытуемый может различными рассуждениями, но наиболее эффективно и наглядно будет рассмотреть фигуры, вписанные в окружность.

Еще одним преимуществом ЕТЗУО по сравнению с другими ЕТЗ является возможность дифференцировать оценки в зависи-

мости от дидактических целей теста и глубины контроля учебных достижений.

Допустим, в предыдущем примере указан ответ 12345. С формальной точки зрения эта комбинация цифр довольно близка к верной, лишь одна пара цифр не на «своих» местах. С другой стороны, испытуемый не знает первой из лежащих в основе правильного ответа закономерностей. Он мог набрать комбинацию цифр наугад, что допускает штрафные санкции, или попытаться рассчитать интересующие его величины на реальных примерах и ошибиться, допустим, в расчете площади правильного шестиугольника. Тогда санкции в его отношении должны быть минимальными. Рассмотрим следующую оценочную шкалу:

1) 1 балл за правильную первую цифру вводимого числа;
2) 0,4 балла за правильную вторую цифру;
3) 0,3 балла за правильную третью цифру;
4) 0,2 балла за правильную четвертую цифру;
5) 0,1 балла за правильную пятую цифру.

Таким образом, если за каждую неверную позицию цифры во вводимом числе начислять ноль баллов, ответ будет оценен в 1,3 балла при максимальном балле 2,0.

Если за каждую неверную позицию цифры в рассмотренном примере вычитать соответствующее этой позиции количество баллов, то оценка за выполнение ЕТЗ составит: $1 – 0{,}4 – 0{,}3 + 0{,}2 + 0{,}1 = 0{,}6$.

ЕТЗУО, в отличие от других рассмотренных видов ЕТЗ (с выбором единственно верного ответа, с множественным выбором верных ответов, ЕТЗУС, ЕТЗУП), предоставляет разработчику больше возможностей в формировании оценочной шкалы и одновременно обеспечивает его более интересными для анализа результатами контроля. Проиллюстрируем это утверждение следующим примером.

«Столицей штата Аляска (США) является:

1. Денвер.
2. Анкоридж.
3. Дюнкерк.
4. Джуно.
5. Зурбаган».

Правильная комбинация цифр – 42135. В результате тестирования было получено N = 32 ответа. В таблице 4.1 приведены значения отклонений:

$$\Delta x_i = x_k - x_{ki};\ k,\ i = 1, \ldots, 5,$$

где k – правильный номер позиции города (цифра последовательности), ki – номер позиции текущего ответа и количества таких отклонений.

Таблица 4.1

Величины и количества отклонений Δx_i от правильных позиций городов в ответах тестируемых

Город	+4	+3	+2	+1	0	–1	–2	–3	–4
Джуно	–	–	–	–	4	3	7	8	10
Анкоридж	–	–	–	17	8	6	1	0	–
Денвер	–	–	6	12	10	4	0	–	–
Дюнкерк	–	4	5	5	13	5	–	–	–
Зурбаган	1	4	4	6	17	–	–	–	–

Из первичных аналитических данных могут быть рассчитаны количественные характеристики, позволяющие в совокупности с соответствующими критериями оценить успешность решения теста [89], например среднее отклонение по каждой k-й позиции:

$$\Delta X_k = \frac{1}{N} \sum_{i=-4}^{4} \Delta x_i \cdot n_k(\Delta x_i), \qquad (4.1)$$

где $n_k(\Delta x_i)$ – количество отклонений величиной Δx_i, и дисперсия отклонений:

$$D_k = \frac{1}{N} \sum_{i=-4}^{4} (\Delta x_i)^2 \cdot n_k(\Delta x_i). \qquad (4.2)$$

Результаты вычислений приведены в табл. 4.2.

Таблица 4.2

Статистические характеристики решения ЕТЗУО

Город	$n_k(0)$	$n_k(0)/N \cdot 100\%$	ΔX_k	D_k
Джуно	4	12,5%	- 2,53	8,21
Анкоридж	8	25%	0,313	0,84
Денвер	10	31,3%	0,625	1,25
Дюнкерк	13	40,6%	0,688	2,06
Зурбаган	17	53,1%	0,938	2,31
среднее значение	10,4	32,5%	1,02*	2,93

* – усреднены абсолютные значения

Во второй графе табл. 4.2 представлено количество верно указанных в ответах испытуемых позиций города. Первую позицию правильно отметили всего четыре из тридцати двух отвечавших. Полностью правильный ответ на ЕТЗ дали два испытуемых. Большинство отвечавших назвали столицей штата Аляска его крупнейший город Анкоридж (табл. 4.1), Денвер назван столицей штата в шести ответах. Вымышленный город Зурбаган оказался на последнем месте в комбинации цифр более чем в половине ответов. В целом отвечающие показали слабое знание вопроса, так как правильная позиция характеризуется максимальной дисперсией отклонений, в 2,8 раза превышающей среднее значение данной величины. Лишь одной позиции (2 – Анкоридж) соответствует среднеквадратичное отклонение (СКО) $\sigma_k = \sqrt{D_k}$ меньше единицы, усредненное для всех ответов значение СКО $\overline{\sigma}$ =1,71, а СКО правильного ответа $\sigma_1 = 2{,}87$, т. е. усредненная величина промаха от минимально необходимого знания составила почти три балла из четырех возможных. Очевидно, для ЕТЗУО свидетельством наличия минимально необходимого знания у испытуемых следует считать результат: $\overline{\sigma} > \sigma_1$, однако в рассматриваемом случае СКО правильного ответа в 1,68 раза превышает $\overline{\sigma}$. Таким образом, на значениях и количестве отклонений Δx_i от правильных позиций вариантов ответа в ЕТЗУО, которые сравнительно просто определяются при автоматизированной обработке результатов тестирования, может строиться обобщенный статистический анализ контроля знаний.

При изучении темы «Контроль знаний и умений с применением компьютерных тестов» в программе повышения квалификации профессорско-преподавательского состава университета «Инновационные образовательные технологии» обучающимся было предложено оценить по пятибалльной шкале некоторые качества различных видов ЕТЗ. Усредненные оценки приведены в табл. 4.3.

Таблица 4.3

Оценки качеств тестов разработчиками

Вид ЕТЗ	Характеристики теста					Сумма баллов
	Достоверность оценок	Устойчивость к деградации	Обучающий потенциал	Простота создания	Универсальность применения	
С единственным выбором	3,1	3,3	3,0	4,4	4,2	18,0
С множественным выбором	3,6	3,8	3,2	3,9	3,7	18,2
На установление соответствия	3,8	3,7	3,6	3,9	4,1	19,1
На установление последовательности	3,9	3,6	3,8	3,3	3,6	18,2
С упорядочиванием ответов	4,6	4,3	3,9	3,1	3,8	19,7

ЕТЗУО набрало максимальную сумму баллов, главным образом за счет высокой, по мнению респондентов, достоверности оценок знаний обучающихся, получаемых в тестах на его основе. Кроме того, на первое место среди рассматриваемых видов ЕТЗ опрошенные поставили ЕТЗУО по таким показателям, как «Устойчивость к деградации» (обеспечение достоверности оценок при многократном применении теста) и «Обучающий потенциал», т. е. возможность для составителя теста формировать у обучающихся знания в процессе их тестирования. Очевидно ожидаемо последнее место ЕТЗУО заняло по простоте (трудоемкости) создания. Респонденты отметили, что разработка тестов на основе ЕТЗУО требует от преподавателей не только глубоких познаний в своем предмете, но и творческих способностей и общего кругозора.

Отдельные ЕТЗ на установление последовательности формально очень близки к ЕТЗУО, например: «Распределите цвета видимого спектра электромагнитного излучения по возрастанию длины волны: 1. Красный. 2. Голубой. 3. Зеленый. 4. Желтый. 5. Фиолетовый» Казалось бы, алгоритм решения этого примера является типичным для ЕТЗУО: сначала находится цвет с минимальной длиной волны (красный), затем выстраивается вся последовательность: 14325. Однако для сравнения приведем ЕТЗУО с аналогичным заданием.

«Средняя частота какого диапазона цвета из видимого спектра электромагнитного излучения наиболее близка к средней частоте желтого цвета? Варианты ответа: 1. Красный. 2. Зеленый. 3. Оранжевый. 4. Инфракрасный. 5. Голубой». Правильная последовательность вариантов ответа: 32514 [90]. Очевидно, решение должно начинаться с постановки на пятую позицию инфракрасного диапазона хотя бы по формальному признаку его отсутствия в видимом диапазоне электромагнитных волн или из знания, что ширина инфракрасного диапазона намного превышает ширины спектров всех видимых цветов, вместе взятых. Дальнейшая расстановка позиций основывается на понимании сложности определения границ видимого излучения ввиду субъективизма цветного зрения, приводящей к относительно большой ширине крайних (красного и фиолетового) диапазонов, а также знания близости желтого и оранжевого (так же, кстати, как и синего и голубого) спектров, например из знакомства с иными (с количеством цветов, не равным семи) цветовыми представлениями видимого электромагнитного излучения или запоминания ширины диапазонов основных видимых цветов.

4.2. Моделирование надежности компьютерного теста с повышенной устойчивостью к возникновению ошибок второго рода

Выше нами была рассмотрена обладающая высоким дидактическим потенциалом структура ЕТЗ, решение которого представляет собой последовательность из порядковых номеров предлагаемых вариантов ответа, которую необходимо сформировать

по степени убывания их правильности [92]. По нашему предположению, помимо дополнительной обучающей функции и расширенных возможностей детализированного анализа выявленных знаний, тест, составленный из ЕТЗУО, обладает и повышенной устойчивостью к ошибкам второго рода, т. е. потенциально более надежен по сравнению с КТОЗУО на основе других типов ЕТЗ.

Пример вопроса. Какая планета Солнечной системы наиболее удалена от Солнца? Варианты ответа: 1. Оберон. 2. Меркурий; 3. Уран. 4. Орион. 5. Нептун. Правильная последовательность, вводимая тестируемым с клавиатуры: 53214. Объяснение ответа: 5, Нептун – абсолютно и единственно верный ответ; 3, Уран – 7-я планета Солнечной системы, наиболее удаленная от Солнца, не считая Урана; 2, Меркурий – единственная из оставшихся трех наименований планета Солнечной системы, пусть и ближайшая к Солнцу; 1, 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, не гарантирует положительной оценки за решение задания, а определить расположение ответов по порядку методом проб и ошибок без достаточных знаний предмета потребует большого количества попыток. Если у преподавателя возникнут основания полагать, что КТОЗУО за время использования подвергся деградации настолько, что не обеспечивает объективность оценки за счет увеличения количества ошибок второго рода [47, 61], он может легко восстановить работоспособность теста, изменив некоторые варианты ответа. Так, для нашего примера измененная последовательность вариантов ответа может выглядеть следующим образом: 1. Титан. 2 Сатурн. 3. Уран. 4. Плутон. 5. Нептун. В целях наглядности примера правильная последовательность, вводимая при ответе на ЕТЗ, не изменится – 53214, однако рассуждения, приводящие к ней, будут заметно отличать-

ся от приведенных выше и потребуют новых знаний в устройстве Солнечной системы.

В ЕТЗ с выбором единственно верного ответа или множественным выбором верных ответов замена одних дистракторов на другие практически не снижает вероятность ошибки второго рода в тесте, подвергшемся «взлому», и для повышения надежности ЕТЗ требуется замена и верных вариантов ответа, возможно известных не обладающим реальными знаниями обучающимся, т. е. полная переработка задания. В рассмотренном примере ЕТЗУО сама усложненная структура ответа обеспечивает относительную долговечность задания за счет регулярного редактирования перечня вариантов ответа.

Оценим вероятность успешного прохождения обучающимися КТОЗУО по аналогии с [65, 77, 91] для различных моделей тестируемых в соответствии с предполагаемыми уровнями их учебных достижений (табл. 4.4).

Таблица 4.4

Модели подготовленности тестируемых

Уровень подготовки	Процент известных ответов по подготовке ЕГЭ				
	1	2	3	4	5
0	0	0	0	0	0
1	20	0	0	0	20
2	20	0	0	0	80
3	50	0	0	0	50
4	100	0	0	0	50
5	100	0	0	0	100
6	60	10	10	10	50
7	80	10	10	10	50
8	30	30	30	30	90
9	40	40	40	40	100
10	50	50	50	50	100
11	60	60	60	60	100
12	80	80	60	60	100

Количественные показатели моделей тестируемых определены нами из следующих соображений. Уровни подготовки с 0 по 5 соотносятся с обучающимися, не освоившими соответству-

ющую учебную дисциплину (модуль, раздел, тему) и рассчитывающих исключительно на подсказки и угадывание верных ответов. Уровень 0 представляет интерес для сравнения значений вероятности успешного прохождения теста для КТОЗУО, представленных различными видами ЕТЗ. Количество известных абсолютно верных ответов на вопросы ЕТЗУО (первая позиция вводимой последовательности) определяется для уровней 1–5 в основном степенью деградации теста, а количество известных пятых позиций искомой цифровой комбинации – как степенью деградации текста методом исключения заведомо неверных ответов, так и собственной общей эрудицией тестируемых. Уровни подготовки 6 и 7 соответствуют начальному уровню освоения обучающимися проверяемого учебного материала, на который накладывается подсказка верных ответов, обеспечивающая правильный ввод первой позиции цифровой комбинации. Уровни 8–12 характеризуют частичное, но добросовестное освоение тестируемыми учебного материала, и это знание, начиная с 9-го уровня, позволяет исключать из рассмотрения заведомо неверные ответы.

Рассмотрим алгоритм расчета итогового балла при решении теста, составленного из десяти ЕТЗУО, содержащих каждое по 5 вариантов ответа на вопрос (задание).

1. Выбор правила начисления баллов (табл. 4.5) и условия успешного решения теста.

Таблица 4.5

Начисляемые баллы за решение ЕТЗУО

Позиция во вводимой последовательности цифр	Правильная цифра	Неправильная цифра
1	1,0	0
2	0,4	0
3	0,3	0
4	0,2	0
5	0,1	0

Балл за решение *i*-го ЕТЗУО:

$$\mathrm{B}_i = \sum_{j=1}^{n_i} b_{ij} \cdot \delta_{(xX)ij}, \qquad (4.3)$$

где b_{ij} – значение веса j-й позиции вводимой последовательности (приведены во втором столбце таблицы 2), $\delta_{(xX)ij}$ – символ Кронекера [15] для расположенных на j-й позиции *i*-го ЕТЗУО цифровых значений вводимой (x) и эталонной (X) последовательностей, $n_i = 5$ – количество позиций в последовательности.

С учетом заданных значений b_{ij} максимальный балл за решение теста:

$$B_{max} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} b_{ij} = 20. \qquad (4.4)$$

Тогда условие положительной оценки за решение КТОЗУО:

$$T \geq T_{min0} = 0{,}8 \cdot B_{max} = 16. \qquad (4.5)$$

2. Определение вероятности правильного указания позиции ответа для одного ЕТЗ:

а) определение всех возможных комбинаций цифр вводимых последовательностей при их угадывании с учетом знаний тестируемым некоторого количества правильных ответов: процент знания позиций с номерами 3 и 4 (см. табл. 4.4) означает решение такого же процента ЕТЗУО целиком или исключение соответствующего количества вопросов теста из расчета вероятности угадывания ответов; известные позиции с номерами 1 и 5 (для уровня подготовки 12 – дополнительно позиции с номером 2), не задействованные в известных тестируемым правильных решениях ЕТЗУО, распределяются по заданиям теста случайным образом;

б) расчет вероятности ρ_{ij} угадывания каждой позиции ЕТЗУО с учетом известных для отдельных заданий позиций 1, 2 (для уровня подготовки 11) и 5.

3. Расчет среднего значения вероятности $\overline{\rho}_j$ угадывания верной j-й позиции во всех заданиях КТОЗУО, среднего набираемого балла по каждой позиции и среднего суммарного балла.

4. Вычисление вероятности прохождения КТОЗУО:

а) расчет вероятности генерирования последовательностей цифровых комбинаций при решении теста с учетом количественных показателей уровня подготовки тестируемого;

б) отбор последовательностей комбинаций, гарантирующих выполнение условия (3);

в) расчет итогового значения вероятности выполнения условия (3) как среднего арифметического значения вероятностей отобранных последовательностей комбинаций.

5. Расчет прогнозируемого итогового балла как суммы произведений вероятностей, рассчитанных по формуле Бернулли [46], на соответствующие им баллы.

Результаты расчетов представлены в табл. 4.6.

Таблица 4.6

Результаты моделирования решения теста из десяти ЕТЗУО

Уровень подготовки	Гарантированный балл	Недостающий балл	Балл за счет угадывания	Прогнозируемый итоговый балл	Вероятность положительной оценки
0	0	16	0,167	0,167	$4{,}57\cdot10^{-53}$
1	2,2	13,8	0,288	2,488	$1{,}87\cdot10^{-36}$
2	2,8	13,2	0,763	3,563	$6{,}63\cdot10^{-26}$
3	5,5	10,5	0,719	6,219	$3{,}47\cdot10^{-22}$
4	10,5	5,5	0,958	11,458	$8{,}06\cdot10^{-11}$
5	11	5	1,5	12,5	$2{,}13\cdot10^{-6}$
6	7,4	8,6	1,125	8,525	$2{,}03\cdot10^{-13}$
7	9,4	6,6	1,239	10,639	$2{,}15\cdot10^{-10}$
8	6,6	9,4	1,318	7,918	$2{,}12\cdot10^{-8}$
9	8,6	7,4	2,804	11,404	$2{,}81\cdot10^{-4}$
10	10,5	5,5	3,070	13,570	$2{,}90\cdot10^{-2}$
11	12,4	3,6	3,380	15,780	0,438
12	15,2	0,8	2,431	17,631	0,986

Результаты моделирования показывают следующее:

– тест, составленный из десяти ЕТЗУО, обеспечивает меньшее количество набираемых за счет угадывания баллов (в среднем не более 10 % от B_{max}) и меньшую вероятность получения удовлетворительной оценки тестируемыми с гарантированным баллом $T < T_{min}$ по сравнению с тестами из двадцати ЕТЗ любых других типов [67, 77, 91], что позволяет формировать на основе ЕТЗУО компактные тесты повышенной надежности;

– структура ЕТЗУО препятствует получению незаслуженной удовлетворительной оценки даже тестируемым с изначально высоким гарантированным баллом (уровни 4, 5, 7, 10, 11), если их реальные знания не обеспечивают решение более 60 % заданий теста;

– недостатком алгоритма формирования итогового балла, основанного на выражении (4.3), является независимое и безусловное начисление парциальных баллов за верное указание отдельных позиций вводимых последовательностей, что, во-первых, при анализе обобщенных результатов тестирования не дает реальной картины освоения обучающимися отдельных дидактических единиц содержания учебной дисциплины, во-вторых, не исключает в полной мере побудительных мотивов к «взлому» теста.

В значительной степени указанного недостатка лишен способ подсчета баллов за выполнение ЕТЗ, предусматривающий штрафы за неверные ответы, например:

$$\mathrm{B}_i = \sum_{j=1}^{n_i} \left[b_{ij} \cdot \delta_{(xX)ij} + \left(\delta_{(xX)ij} - 1 \right) \cdot b_{ij} \right]. \qquad (4.6)$$

В данном случае за совпадение позиций во вводимой и эталонной последовательностях ($x = X$) начисляется парциальный балл b_{ij} (см. табл. 4.5), а за несовпадение ($x \neq X$) вычитается аналогичное число. Как следствие, при верно указанной первой цифре вводимой последовательности и неверно введенных остальных цифрах $\mathrm{B}_i = 0$. Следует предположить, что начисление баллов на основе выражения (4.6) сделает невыполнимым условие (4.5) для любого рассматриваемого уровня подготовленности тестируемых. Для повышения вероятности прохождения теста при неизменных (для корректности вычислительного эксперимента) ха-

рактеристиках подготовленности тестируемых изменим условие положительной оценки:

$$T \geq T_{min1} = 0{,}4 \cdot B_{max} = 8; \quad (4.5а)$$

$$T \geq T_{min2} = 0{,}6 \cdot B_{max} = 12; \quad (4.5б)$$

Результаты численного моделирования работы алгоритма с начислением штрафных баллов за неверные ответы для уровней подготовленности тестируемых с наибольшими гарантированными баллами приведены в табл. 4.7.

Таблица 4.7

Результаты моделирования решения теста со штрафами за неверные ответы

Уровень подготовки	Гарантированный балл	Недостающий балл по (3а)	Недостающий балл по (3б)	Балл за счет угадывания с учетом штрафов	Прогнозируемый итоговый балл	Вероятность положительной оценки по (3а)	Вероятность положительной оценки по (3б)
4	10,5	–2,5	1,5	–7,58	2,92	$5{,}59 \cdot 10^{-5}$	$8{,}41 \cdot 10^{-11}$
5	11	–3	1	–6,0	5,0	0,0176	$2{,}1 \cdot 10^{-6}$
7	9,4	–1,4	2,6	–8,24	1,16	$6{,}35 \cdot 10^{-5}$	$9{,}98 \cdot 10^{-10}$
8	6,6	1,4	5,4	–9,23	–2,63	$3{,}23 \cdot 10^{-5}$	$2{,}12 \cdot 10^{-8}$
9	8,6	–0,6	3,4	–5,79	2,81	0,0220	$2{,}51 \cdot 10^{-4}$
10	10,5	–2,5	1,5	–3,27	7,23	0,360	0,0248
11	12,4	–4,4	–0,4	–1,71	10,69	0,947	0,411
12	15,2	–7,2	–3,2	–0,06	15,14	1,0	0,986

Как показывают расчеты, даже изначальное превышение гарантированным баллом величины T_{min} при начислении штрафов за неверные ответы не всегда обеспечивает итоговую положительную оценку. По сравнению со знанием верных ответов в общем (позиция 1 искомой последовательности) 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 освоивших учебный материал обучающихся и проблемные для освоения дидактические единицы содержания учебной дисциплины.

Апробация тестов, составленных из ЕТЗУО, выявила относительно высокий уровень (до 6 % ответов) ошибок первого рода, обусловленных ошибками тестируемых в наборе цифровых комбинаций. Для уменьшения количества таких ошибок требуется разработка более сложных интерфейсов тестовых программных оболочек, которые, например, по введенным цифровым комбинациям визуализируют последовательность ответов перед ее окончательным утверждением, обеспечивают возврат к предыдущему или исходному этапу решения задания (отмену выбора) и др.

Алгоритм оценивания ЕТЗУО с начислением штрафных баллов обладает большей устойчивостью к «взлому» баз знаний и угадыванию верных ответов. В развитии данного алгоритма следует учесть при начислении штрафа степень отклонения цифр в ответе тестируемого на ЕТЗУО от их истинных значений, т. е. корреляцию вводимой и эталонной последовательностей [92]. Наиболее целесообразен учет отклонений для первой и последней позиций последовательности. Общий вид выражения для начисления штрафных баллов за неверное решение i-го ЕТЗУО может иметь вид:

$$\mathrm{R}_i = \sum_{j=1}^{n_i} r_{ij} = \sum_{j=1}^{n_i} \left(\varphi_j + |x - X|_{ij} \cdot \gamma_j\right), \qquad (4.7)$$

где r_{ij} – штраф за неверное указание j-й позиции i-го ответа из их общего количества n_i, $|x - X|_{ij}$ – модуль разности указанного и верного номеров j-й позиции i-го ответа, φ_j и γ_j – величины, определяющие значимость j-й позиции эталона.

Тогда общая формула расчета балла за решение i-го ЕТЗУО:

$$\mathrm{B}_i = \sum_{j=1}^{n_i} b_{ij} \cdot \delta_{(xX)ij} - \left(\varphi_j + |x - X|_{ij} \cdot \gamma_j\right). \qquad (4.8)$$

Алгоритм расчета B_i с начислением штрафных баллов, по нашему мнению, особенно востребован на экзамене как обеспечивающий объективную дифференциацию положительных оценок или на завершающем этапе тестирования, которому предшествует тест без начисления штрафа или на основе ЕТЗ с более простой структурой.

Применение алгоритмов расчета итоговых баллов за решение КТОЗУО с начислением штрафов за неверные ответы обеспечивает возможность адаптивной подстройки тестов [84, 86, 87] под тестируемый контингент путем изменения параметров φ_j, γ_j и T_{min}.

Минимизация ошибок второго рода, обеспечиваемая структурой ЕТЗУО, позволяет формировать высоконадежные КТОЗУО, вдвое компактнее по сравнению с тестами на основе ЕТЗ других типов, что делает общую трудоемкость разработки тестов на основе ЕТЗУО сравнимой с разработкой КТОЗУО на основе ЕТЗ с более простой структурой.

4.3. Практические рекомендации по обеспечению достоверности оценивания учебных достижений обучающихся компьютерными тестами

Помимо количества и структуры ЕТЗ, на вероятность ошибок первого и второго рода могут оказывать влияние многие факторы, с которыми необходимо считаться разработчикам тестов. Не рассматривая общих требований, отраженных в 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) [11–13], приведем некоторые рекомендации, основанные на практическом опыте применения тестов для оценивания учебных достижений обучающихся авторами монографии.

1. Принцип синхронизации (учет степени актуальности проверяемых знаний и умений).

Содержание и уровень сложности заданий должны учитывать фактор забывания обучающимися содержания изученного материала и частичной утраты сформированных у них умений [94–97]. В особенности этот фактор влияет на оценки тестируемых на контрольно-проверочных мероприятиях мониторинга качества реализации образовательных программ, проводимого об-

разовательной 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 сложности и объемами проверяемого содержания для дидактических единиц содержания обучения, 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 [60, 71, 98], 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.

2. Принцип однородности заданий.

Как показывает наш опыт разработки и применения тестов, к дополнительным ошибкам первого рода приводит многообразие ЕТЗ, из которых составлен тест, например наличие в нем наряду с ЕТЗ с выбором единственно верного ответа ЕТЗ с множественным выбором или ЕТЗ открытого типа. Несмотря на наличие пояснений (инструкций) или внешние отличия структуры ЕТЗ тестируемым требуется время на «переключение» мыслительных стереотипов, возрастает вероятность ошибок из-за невнимательности, несвоевременного прочтения или недопонимания инструкций. Дополнительным условием правильного и быстрого восприятия ЕТЗ тестируемым является стилевая однородность формулировок заданий и вариантов ответов.

Приведем пример стилевой неоднородности формулировок ответов.

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 использовать ЕТЗ различных видов следует делить тест на блоки, в каждом из которых представлены ЕТЗ одного вида, или разбивать процедуру тестирования на этапы, каждый из которых представляет собой мини-тест, составленный из однородных ЕТЗ.

3. Принцип пирамиды «сложность – количество тестируемых».

Совершенно необязательно использовать большие ресурсы КТОЗУО на проведение контрольно-оценочных мероприятий, разных по уровню сложности задач контроля. Во многом уровень сложности задач по оцениванию знаний и умений обучающихся определяется уровнем мотивации и образовательных амбиций тестируемых. Для оценки знаний обучающегося, на протяжении всего периода изучения дисциплины балансирующего между «двойкой» и «тройкой», нет смысла применять тесты с большой дифференцирующей способностью, содержащие задания различных уровней сложности. Если сложность заданий, отобранных педагогом для формирования теста, можно представить четырьмя уровнями: простые, средней сложности, повышенной сложности, сложные, – то для слабоуспевающих обучающихся достаточно включения в тест ЕТЗ первых двух уровней, а ответы на задания повышенной сложности и сложные он будет угадывать, что не скажется существенным образом на его оценке (при адаптивном алгоритме оценивания), но может привести к снижению надежности теста, когда тестируемый по окончании контроля знаний перескажет запомненные им вопросы теста и варианты ответа на них другим обучающимся. Принципы возрастающей трудности и соответствия содержания теста целям тестирования, сформулированные В.С. Аванесовым [12, 13], 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 сложности применяются к оценке знаний всех обучающихся, более сложные тесты – к оценке знаний тех из них, кто показал высокие результаты на предыдущем уровне. Данный подход позволяет модернизировать базы ЕТЗ частично, в основ-

ном в отношении сравнительно простых заданий, а также сокращать общее время тестирования и последующей переэкзаменовки.

4. Принцип исключения тривиально правильных ответов.

Далеко не всегда удается сформулировать ЕТЗ таким образом, чтобы оно не подразумевало тривиальный выбор правильного ответа, являющегося стандартизированным определением какого-либо понятия или термина, или наоборот. Однако к исключению такого рода ситуаций и формулированию ЕТЗ 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, необходимо стремиться. Выбор 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.

Пример 1. 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 2. 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, удается не всегда, а время попыток поиска в Интернете или получения иных

подсказок со стороны ограничено, поэтому вероятность возникновения ошибки второго рода в ответе на ЕТЗ с нетривиальной формулировкой задания существенно уменьшается, а обучающийся, поставленный в условия нетривиальных формулировок ЕТЗ и ответов на них (пример 3) вынужден будет считаться с необходимостью более глубокого погружения в изучаемый материал.

Пример 3. 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 ЕТЗ, не должны быть в одинаковой степени похожи на правильный ответ, их «похожесть» на правильный ответ следует дифференцировать. Приведем пример неудачного, по нашему мнению, подбора дистракторов.

Пример 4. 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 ЕТЗ не содержит полностью неверный ответ – все они являются частично правильными. Поэтому, анализируя выбор тестируемым ответа, сделанного осознанно или наугад, преподаватель не сможет определить степень незнания предмета, что наряду с собственно оцениванием является одной из целей тестирования учебных достижений обучающихся. В этом отношении ЕТЗ из примера 1 выглядит намного предпочтительнее. Выбор тестируемым варианта «г» из примера 1 свидетельствовал бы о полном непонимании им сущности и механизмов проявления электрического тока; выбор варианта «в» – скорее о попытке угадать ответ, основываясь на «наукообразии» предлагаемого варианта, или о наличии обрывочных несистем-

ных 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 адекватности времени решения виду задания.

Разработчики тестов нередко пренебрегают дифференцированием времени ответа на ЕТЗ в зависимости от сложности и особенно структуры последнего. В то же время вполне умозрительным представляется вывод о том, что на решение ЕТЗ с множественным выбором верных ответов требуется больше времени, чем на решение ЕТЗ с выбором единственно верного ответа при условии одинаковой сложности обоих ЕТЗ. Решение ЕТЗУП как минимум требует столько же времени, что и решение ЕТЗ с множественным выбором верных ответов, а ЕТЗУО – однозначно больше.

Задание времени ответа на ЕТЗ различных видов в КТОЗУО может быть определено на основе эмпирических данных.

Важно минимизировать время ответа на ЕТЗ с простой структурой, ориентированного на проверку базовых или элементарных знаний. Автоматизм решения таких ЕТЗ косвенно свидетельствует о системности и высоком уровне проверяемых знаний.

7. Принцип редактируемости заданий.

Выбор структуры ЕТЗ и формулирование вопросов (заданий) и ответов на них целесообразно осуществлять таким образом, чтобы их можно было видоизменять (редактировать вопросы (задания), менять или редактировать дистракторы) с минимальными трудозатратами. В этом заключается «ремонтопригодность» (по аналогии с техническими объектами) тестов, являющаяся одним из показателей надежности. Регулярное редактирование ЕТЗ вместо их замены позволяет, во-первых, снижать трудоемкость процесса, направленного на уменьшение вероятности возникновения ошибок второго рода, во-вторых, оставляя без изменения тестируемое содержание, не дает оснований пересмотра содержательной валидности теста, в-третьих, прецизионно регу-

лировать сложность теста, уменьшая или увеличивая ее для одних ЕТЗ и оставляя неизменной для других.

8. Принцип контрольных «закладок».

В своей практике разработки тестов мы используем ЕТЗ сверхсложного уровня или с содержанием, выходящим за пределы изучаемого материала. Если процент решения таких ЕТЗ превышает вероятность случайного угадывания, это свидетельствует о том, что тест подвергся «взлому», и база ЕТЗ подлежит замене или редактированию. В разделе 3.1 механизм автоматизированной проверки теста на «взлом» реализован на основе сравнения процента решения простых и сложных заданий. Творческий подход к конструированию КТОЗУО обеспечивает широкое разнообразие подобного рода приемов – индикаторов повышения вероятности отказа теста.

Заключение

Любой разработчик тестов для контроля знаний и умений обучающихся рано или поздно сталкивается с проблемой обеспечения надежности контрольно-измерительных материалов как в аспекте долговечности, так и минимально достижимой вероятности ошибок оценивания. На этапе проектирования теста наиболее актуальны математические или экспертные модели, обеспечивающие достоверность априорных оценок надежности тестов и тем самым позволяющие прогнозировать степень объективности результатов тестирования и обосновать корректность и оптимальность решений разработчиков тестов, направленных на минимизацию ошибок оценивания учебных достижений испытуемых.

Ряд авторов основывают определение надежности теста на воспроизводимости его результатов (оценок), что оставляет за пределами анализа систематическую погрешность оценок, например из-за неправильного соотнесения базы заданий и базы знаний. Нами предложено определение надежности теста из его сопоставления с сортировочным или отбраковочным автоматом, базирующееся на общих положениях теории надежности, и в основу критерия надежности теста положено его безошибочное оценивание знаний и умений обучающихся. В связи с этим надежность теста характеризуется вероятностью его полного или частичного отказа, которая, в свою очередь, напрямую зависит от ошибок первого и второго рода, приводящих к занижению и завышению оценок соответственно. Если ошибки первого рода носят методический или технический характер и могут быть устранены в процессе отладки и эксплуатации теста, то в ошибках второго рода преобладает деградационная составляющая, увеличивающая частоту возникновения ошибочных оценок с увеличением времени эксплуатации теста и обусловленная запоминанием обучающимися правильных ответов и передачей этой информации от одних испытуемых другим.

Такой подход не только позволил по-новому, через вычисление вероятности ошибки второго рода, обусловленной частичным угадыванием правильных ответов, оценить качество тестов, создаваемых на основе различных алгоритмов и видов тестовых заданий, но и предложить решения, активизирующие творчество педагогов и обеспечивающие высокую достоверность результатов тестирования при компактности баз тестовых заданий и их выборок для оценивания учебных достижений.

Литература

1. Ефремова Н.Ф. Тестовый контроль в образовании: учеб. пособие. М.: Логос, 2005.
2. Векслер В.А. Возникновение тестологии // Современные научные исследования и инновации. 2015. № 5. Ч. 4 [Электронный ресурс]. URL: http://web.snauka.ru/issues/2015/05/48321 (дата обращения: 14.09.2020).
3. РМГ 29-2013 ГСИ. Метрология. Основные термины и определения.
4. ГОСТ 16263-70. Государственная система обеспечения единства измерений.
5. РМГ 29-99 ГСИ. Метрология. Основные термины и определения.
6. Звонников В.И., Найденова Н.Н., Никифоров С.В., Челышкова М.Б. Шкалирование и выравнивание результатов педагогических измерений. М.: Логос, 2003.
7. Балыхина Т.М. Словарь терминов и понятий тестологии. М.: РУДН, 2000.
8. 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.
9. Аванесов В.С. Форма тестовых заданий. М.: Центр тестирования, 2005.
10. Хлебников В.А., Михалева Т.Г. Отраслевой стандарт, педагогические тесты, термины и определения. М.: Центр тестирования, 2001.
11. Аванесов В.С. Композиция тестовых заданий. М.: АДЕПТ, 1998.
12. Аванесов В.С. Основы педагогической теории измерений // Педагогические измерения. 2004. № 1. С. 15–21.
13. Аванесов В.С. Теория и практика педагогических измерений (материалы публикаций в открытых источниках и Интернете). Екатеринбург: ЦТ и МКО УГТУ-УПИ, 2005.

14. Челышкова М.Б. Теоретико–методологические и технологические основы адаптивного тестирования в образовании: дис. … д-ра пед. наук. М., 2001.

15. Летова Л.В. Точность моделирования латентных переменных с помощью модели Раша (часть 1) // Современные научные исследования и инновации. 2014. № 6. Ч. 1 [Электронный ресурс]. URL: http://web.snauka.ru/issues/2014/06/34399 (дата обращения: 24.11.2020).

16. Летова Л.В. Исследование качества теста как измерительного инструмента // Дистанционное и виртуальное обучение. 2013. № 11.

17. Попов А.П. Критический анализ параметрических моделей Раша и Бирнбаума // Инновационные методы и средства оценки качества образования: материалы IV Всерос. науч.-метод. конф. / Московский государственный университет печати; Независимый центр тестирования качества обучения, 20–21 апр. 2006 г. М.: Изд-во МГУП, 2006.

18. Челышкова М.Б., Ковалева Г.С. Основные подходы к оценке качества подготовки обучаемых в России и за рубежом // Квалиметрия человека и образование: методология и практика. Восьмой симпозиум. М.: ИЦПКПС, 1998.

19. Высшее образование в XXI веке. Подходы и практические меры. Всемирная конференция по высшему образованию ЮНЕСКО. Париж, 1998.

20. Челышкова М.Б. Теория и практика конструирования педагогических тестов: учеб. пособие. М.: Логос, 2002.

21. Стоунс Э. Психопедагогика. М., 1984.

22. Татур А.О. Тесты в учебном процессе // Новые технологии в обучении и контроле знаний учащихся: материалы науч.-практ. конф. М., 1999.

23. Шмелев А.Г., Бельцер А.И. и др. Перспективы компьютерного тестирования: валидность и надежность «Телетестинга» // Развитие системы тестирования в России: тез. докл. Всерос. конф. М., 1999. Ч. 3.

24. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей, 2000.

25. Энциклопедия психологических тестов. Личность. Мотивация. Потребность / под ред. А. Карелина. М.: АСТ, 1997.

26. Анастази А., Урбина С. Психологическое тестирование. 7-е междунар. изд. СПб.: Питер, 2001.

27. 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 симпозиум. М.: ИЦПКПС, 2002. Ч. 3.

28. Кларин М.В. Инновационные модели обучения в зарубежных педагогических поисках. М.: Арена, 1994.

29. Переверзев В.Ю. Технология разработки тестовых заданий: справ. руководство. М.: Е-Медиа, 2005.

30. Старостенко И.Н., Хромых А.А Элементы статистической обработки данных: учеб. пособие. Краснодар: Краснодарский университет МВД России, 2020.

31. Татарова Г.Г. Методология анализа данных в социологии (введение): учеб. для вузов. М.: NOTA BENE, 1999.

32. Кислицына Е.А. Фонетические тесты как средство диагностики и контроля при обучении русскому языку китайских учащихся-нефилологов (начальный этап): дис. … канд. пед. наук. СПб., 1995.

33. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов: учеб. / пер. с англ. Н.Н. Найденовой, В.Н. Симкина, М.Б. Челышковой; под общ. ред. В.И. Звонникова, М.Б. Челышковой. М.: Логос, 2010.

34. Денисенко Л.Г. Тестовые материалы в условиях реализации ФГОС для учреждений профессионального образования (из опыта разработки). Ч. 1. Новосибирск: Новосибирский институт мониторинга и развития образования, 2014. (Серия «Современные средства оценивания в образовании»).

35. Бочкарева Т.Н. Современные средства оценивания результатов обучения: учеб.-метод. пособие / Елабужский институт Казанского федерального университета. Махачкала: Апробация, 2019.

36. ГОСТ Р ИСО 9000-2015. Системы менеджмента качества. Основные положения и словарь. М.: Стандартинформ, 2015.

37. Аванесов В.С. Современные методы обучения и контроля знаний. М.: ИЦПКПС, 1998.

38. Национальная энциклопедическая служба [Электронный ресурс]. URL: https://vocabulary.ru/termin/validnost.html (дата обращения: 05.11.2020)

39. Рапопорт И.А., Сельг Р., Соттер И. Тесты в обучении иностранным языкам в средней школе. Таллин: Валгус, 1987.

40. Коккота В.А. Лингводидактическое тестирование: науч.-теорет. пособие. М.: Высшая школа, 1989.

41. Allen J.P.B., Devies A.The Edinburgh Course in Applied Linguistics: Testing and experimental methods. London: Oxford University Press, 1977.

42. Анастази А. Психологическое тестирование: в 2 кн. Кн. 1 / пер. с англ. под ред. К.М. Гуревича, В.И. Лубовского. М.: Педагогика, 1982.

43. Смирнова Г.И. Разработка тезауруса педагогических измерений Г. Раша // Педагогические измерения. 2005. № 4. С. 62–64.

44. Мальцев А.В., Наймушина О.Э. Тестология в образовании: вчера, сегодня, завтра // Известия Уральского государственного университета. 2008. № 60. С. 7–14.

45. Пугачев В.П. Тесты, деловые игры, тренинги в управлении персоналом: учеб. для студентов вузов. М.: Аспект Пресс, 2003.

46. Гмурман В.Е. Теория вероятностей и математическая статистика: учебник для прикладного бакалавриата. 12-е изд. М.: Юрайт, 2017.

47. Острейковский В.А. Теория надежности: учеб. для вузов 2-е изд., испр. М.: Высшая школа, 2008.

48. Аванесов В.С. Основные понятия педагогической тестологии // Научные проблемы тестового контроля знаний: тез. докл. участников школы-семинара. М.: ИЦПКПС, 1994. С. 105–108.

49. Попов Д.И., Попова Е.Д. Экспертиза качества тестовых заданий: учеб. пособие. М.: МГУП, 2008.

50. Аванесов В.С. Научные проблемы тестового контроля знаний. М.: ИЦПКПС, 1994.

51. Аванесов В.С., Володин Б.В., Короза В.И. Опыт построения теста для оценки знаний студентов // Научная организация учебного процесса. 1976. Вып. 3, ч. 1.

52. Деменчёнок О.Г. Математические основы Rasch Measurement // Педагогические измерения. 2010. № 1.

53. 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.

54. Аванесов В.С. Применение тестовых форм в Rasch Measurement // Педагогические измерения. 2005. № 4. С. 3–20.

55. Адаптивное тестирование: учеб.-метод. пособие / Н.М. Опарина, Г.Н. Полина, Р.М. Файзулин, И.Г. Шрамкова. Хабаровск: ДВГУПС, 2007.

56. Лопаткина, Е.В. Современные средства оценивания результатов обучения: учеб. пособие / Владим. гос. ун-т имени Александра Григорьевича и Николая Григорьевича Столетовых. Владимир: Изд-во ВлГУ, 2012.

57. Павловская О.О. Статические методы оценки надежности программного обеспечения // Вестник Южно-Уральского государственного университета. Сер.: Компьютерные технологии, управление, радиоэлектроника. 2009. № 26 (159). С. 35–37.

58. Ямпурин Н.П., Баранова А.В. Основы надежности электронных средств: учеб. пособие для студентов вузов / под ред. Н.П. Ямпурина. М.: Академия, 2010.

59. Денисова О.П. Моделирование системы квалиметрии образовательного процесса в вузе // Вестник Гуманитарного института Тольяттинского государственного университета. 2012. № 2. С. 36–39.

60. Гибадуллина Р.Н. Контроль остаточных знаний // Вестник Казанского государственного энергетического университета. 2013. № 4. С. 113–115.

61. Булгаков О.М., Дедикова А.О. О применимости методологического аппарата теории надежности к оценке качества тестов для проверки знаний // Вестник Воронежского института ФСИН России. 2017. № 4. С. 214–222.

62. ГОСТ 27.002-2009. Надежность в технике. Термины и определения. М.: Стандартинформ, 2011.

63. Ким В.С. Тестирование учебных достижений. Уссурийск: Издательство УГПИ, 2007.

64. Бугров Я.С., Никольский С.М. Высшая математика: учеб. для вузов: в 3 т. / под ред. В.А. Садовничего. 6-е изд., стереотип. М.: Дрофа, 2004.

65. 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.

66. Кононова О.В., Садон Е.В., Якимова З.В. Методика оценки сформированности компетенций на уровне учебной дисциплины // Территория новых возможностей. Вестник Владивостокского государственного университета экономики и сервиса. 2013. № 5 (23). С. 76–87.

67. Галимзянов Х.М., Попов Е.А., Сторожева Ю.А. Формирование и оценка компетенций в процессе освоения образовательных программ ФГОС ВО: науч.-метод. пособие. Астрахань, Астраханский ГМУ, 2017.

68. Иванова Л.А. Оценка уровня сформированности общих и профессиональных компетенций с помощью современных педагогических приемов // Молодой ученый. 2016. № 2. С. 799–804.

69. Жевлакович С.С. К вопросу об оценке результатов освоения основных образовательных программ высшего образования при реализации компетентностного подхода к организации образовательного процесса // Международный журнал психологии и педагогики в служебной деятельности. 2018. № 4. С. 20–25.

70. Гривенная Е.Н., Булгаков О.М. Модель оценки уровня сформированности компетенций обучающихся в образовательной организации МВД России // Вестник Краснодарского университета МВД России. 2018. № 2. С. 116–120.

71. Ершиков С.М., Иванова И.В. Мониторинг уровня остаточных знаний студентов медицинского университета // Ярославский педагогический вестник. 2017. № 5. С. 139–144.

72. ГОСТ Р 50779.10-2000. Статистические методы. Вероятность и основы статистики. Термины и определения.

73. Жунусакунова А.Д. Разновидности заданий в тестовой форме // Актуальные вопросы современной педагогики: материалы II Междунар. науч. конф. (г. Уфа, июль 2012 г.). URL https://moluch.ru/conf/ped/archive/60/2572/ (дата обращения: 03.11.2019).

74. Рогова С.И., Калишев М.Г., Жарылкасын Ж.Ж., Жиенбекова А.Ж., Сабиден Г.С. Использование тестов множественного выбора при оценке результатов самостоятельной работы студентов // Медицина и экология. 2016. № 3. С. 142–144.

75. Шмелев А.Г. Компьютеризация экзаменов: проблема защиты от фальсификаций // Сборник материалов конференции ИТО-2001. Секция VI: ИТ в контроле и оценке результатов обучения. 2001.

76. 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.

77. 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 ФСИН России. 2019. № 4. С. 62–69.

78. Рудинский И.Д., Литвинов К.А. Модель автоматизированного оценивания качества тестовых контрольно-измерительных материалов // Ученый записки ИИО РАО. 2012. № 40. С. 74–90.

79. Беспалько В.П. Методы оценки результатов тестирования учащихся / Вопросы тестирования в образовании. 2007. № 2. С. 23–34.

80. Шойтов Д.В. Применение принципов проектирования UML при использовании предметно-критериальной методики проектирования тестов // Ученые записки. Электронный научный журнал Курского государственного университета. 2010. № 1 (3).

81. Дуплик С.В. Основные модели современной теории тестирования // Вопросы тестирования в образовании. 2002. № 7. С. 56–67.

82. Булгаков О.М., Ладыга А.И., Рябошапко О.Н. Обобщенная модель отбора содержания контроля остаточных знаний // Вестник Воронежского института МВД России. 2019. № 2. С. 41–48.

83. URL: https://studbooks.net/1930719/pedagogika/metody_i_modeli_intellektualnogo_avtomatizirovannogo_kontrolya_znaniy (дата обращения: 18.09.2020).

84. Рудинский И.Д. Принципы интеллектуального автоматизированного тестирования знаний // Сборник материалов конференции ИТО-2001. Секция VI: ИТ в контроле и оценке результатов обучения. 2001.

85. Дуплик С.В. Интеллектуальные обучающие и контролирующие системы // Информатика. Информационные технологии. Средства и системы. 2000. № 2. С. 87–90.

86. Сердюков В.И. Особенности адаптированного автоматизированного контроля знаний // Ученый записки ИИО РАО. 2012. № 40. С. 62–73.

87. Зайцева Л.В., Прокофьева И.О. Модели и методы адаптивного контроля знаний // Образовательные технологии и общество. 2004. № 4. Т. 7. С. 265–277.

88. Булгаков О.М. Совершенствование структуры компьютерных тестов // Новые информационные технологии в процессе подготовки современного специалиста: межвузовский сборник научных трудов. Липецк: ЛГПИ, 1999. Вып.2. С. 13–20.

89. Булгаков О.М., Ладыга А.И., Рябошапко О.Н. Интерпретация результатов контроля остаточных знаний с применением элементов корреляционного анализа и математической статистики // Вестник Воронежского института ФСИН России. 2018. № 2. С. 33–37.

90. URL: https://www.sites.google.com/site/sergkraskaa/elektromagnitnye-volny/vidimoe-izlucenie (дата обращения: 18.09.2020).

91. Булгаков О.М., Старостенко И.Н., Хромых А.А., Дедикова А.О. Моделирование надежности тестов с усложненной структурой тестовых заданий // Вестник Воронежского института ФСИН России. 2020. № 2. С. 62–70.

92. 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.

93. Ильин А.В., Позняк Э.Г. Линейная алгебра. М.: Физматлит, 2007.

94. 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: 27.03.2020).

95. Строганов В.Ю., Макаренко Л.Ф., Ярцев М.И., Ягудаев Г.Г. Модели забывания учебной информации для системы подготовки персонала // Материалы IX Международной заочной научно-практической конференции молодых ученых «Теория и практика применения информационных технологий в промышленности и на транспорте» (г. Москва, 12 ноября 2013 г.). auts.esrae.ru

96. Буймов А.Г. Закономерности поведения кривых забывания // Доклады ТУСУРа. Т. 20, № 4, 2017. С. 138–141.

97. Майер Р.В. Имитационное моделирование изучения студентами вузовского курса, учитывающее психологические закономерности усвоения и забывания // Научно-методический электронный журнал «Концепт». 2015. № 12 (декабрь). С. 116–120. URL: http:e-koncept.ru/2015/15430.htm.

98. Денисова О.П. Подготовка студентов к контролю остаточных знаний на основе обобщающего повторения // Вектор науки Тольяттинского государственного университета. 2012. № 4. С. 86–88.

Оглавление

Научное издание

Булгаков Олег Митрофанович
Старостенко Игорь Николаевич
Хромых Анна Алексеевна
Дедикова Анна Олеговна

МОДЕЛИ ОЦЕНКИ КАЧЕСТВА ТЕСТОВ ДЛЯ КОНТРОЛЯ ЗНАНИЙ

Редактор *Т. Г. Кривошеева*
Компьютерная верстка *Г. А. Артемовой*

ISBN 978-5-9266-1728-0

9 785926 617280

Подписано в печать 16.11.2020. Формат 60х84 1/16.
Усл. печ. л. 8,1. Тираж 100 экз. Заказ 97.

Краснодарский университет МВД России.
350005, г. Краснодар, ул. Ярославская, 128.