электронное научно-техническое издание
НАУКА и ОБРАЗОВАНИЕ
Эл № ФС 77 - 30569. Государственная регистрация №0421100025. ISSN 1994-0408

Тестовый метод контроля качества обучения и критерии качества образовательных тестов. Обзор

 # 04, апрель 2011

авторы: Белоус В. В., Домников А. С., Карпенко А. П.

УДК 519.6

МГТУ им. Н.Э. Баумана
karpenko@rk6.bmstu.ru

Введение

Основным средством формального контроля качества обучения является тестирование. В педагогике и психологии тестированием (от англ. test) называется экспериментальный метод, основанный на стандартизированных заданиях, которые позволяют измерить психофизиологические и личностные характеристики, а также знания, умения и навыки испытуемого [1]. В широкой интерпретации термин тестирование включает в себя тестовый метод, результат тестирования и интерпретацию результатов тестирования.

Тестовый метод контроля качества обучения имеет ряд несомненных преимуществам перед другими педагогическими методами контроля: высокая научная обоснованность теста; технологичность; точность измерений; наличие одинаковых для всех испытуемых правил проведения испытаний и правил интерпретации их результатов; хорошая сочетаемость метода с современными образовательными технологиями.

Тесты начали применяться в 1864 году Дж. Фишером в Великобритании для проверки знаний учащихся. Теоретические основы тестирования были разработаны английским психологом Ф. Гальтоном в 1883 году. Термин "тест" впервые ввёл американский психолог Дж. Кеттел в 1890 году. Первый стандартизированный педагогический тест был составлен американским психологом Э. Торнодайком. Американский психолог К. Спирмен разработал основные методы корреляционного анализа для стандартизации тестов и объективного измерения тестологических исследований. Статистические методы Спирмена, основанные на использовании методов факторного анализа, сыграли большую роль в дальнейшем развитии тестирования. Отметим, что развитие тестирования стало одной из основных причин, обусловивших проникновение в психологию и педагогику математических методов.

В России составление и применение тестов относится к 20-м годам прошлого века. Первая серия тестов для школ была опубликована в 1926 году [1].

Можно выделить следующие этапы в эволюции контроля знаний [2].

1) *Традиционный контроль*. Для оценки знаний обучающихся используют такие формы контроля, как контрольная работа, коллоквиум, курсовая работа и т.д. Преподаватель подготавливает соответствующие варианты заданий, проверяет и оценивает результаты работы учащихся.

2) *Контроль с использованием не компьютерных средств*. В этом случае для контроля используют заранее подготовленные бланки, содержащие контрольные задания (тесты). Учащиеся заполняют бланки, решая задания и отвечая на вопросы. Преподаватель проверяет работы, используя специальные трафареты и таблицы ответов.

3) *Контроль с использованием технических устройств*. В данном варианте контроля учащийся, получив от преподавателя индивидуальный набор тестовых заданий, выполняет его и вводит в техническое устройство номер своего варианта и результат решения каждого задания, а устройство проверяет введенные ответы, рассчитывает и выводит оценку за работу.

4) *Компьютерный контроль*. Здесь контроль знаний обеспечивают специальные компьютерные программы, в которых осуществляется формирование, вообще говоря, индивидуального набора тестовых контрольных заданий каждому учащемуся, вывод заданий на экран монитора, анализ ответов учащегося, выставление результирующей оценки, хранение результатов контроля и данных о работе учащегося.

5) *Удаленный контроль***.** Появление данного подхода к контролю знаний обусловлено, прежде всего, широким использованием в учебном процессе возможностей сети *Internet*. Отличительными чертами удаленного контроля знаний является свобода выбора учащимся темпов тестирования, его времени и места.

Тестирование является одним из видов педагогических измерений, к которым относятся также рейтинг и мониторинг. В своей основе мониторинг относится к сфере управления образованием и по отношению к педагогическим измерениям является лишь поставщиком показателей качества образования.

Работа посвящена задаче оценки качества тестов, которую можно считать частью проблемы анализа заданий [3]. Анализ заданий может быть рациональным (оценочным) и эмпирическим (статистическим). Рациональный анализ заданий предполагает неформальный анализ каждого из заданий теста и реализуется вне автоматической обучающей системы (АОС). Эмпирический анализ заданий означает анализ таких характеристик теста, как его надежность, валидность, трудность, дискриминативность и т.д. [4]. В работе рассматриваются критерии, которые соответствуют указанным аспектам качества тестов.

Отметим следующее обстоятельство. Наряду с обсуждением качества тестов, можно говорить о качестве тестовых результатов, полученных с их помощью. В настоящее время отчетливой является тенденция, в соответствие с которой считается более правильным обсуждать именно вопрос не качества тестов, а качества результатов тестирования с их помощью. Мы, однако, в данной работе останемся на классических позициях и не будем обсуждать качество результатов использования тестов.

Современный уровень развития информационных и коммуникационных технологий открывает возможности создания автоматизированных систем тестирования знаний, которые обычно представляют собой соответствующие подсистемы АОС. Методологическую основу таких систем составляют методы математической статистики, теории принятия решений и искусственного интеллекта (в частности, нечеткая логика и теория экспертных оценок), а также новейшие достижения современной педагогической науки. В англоязычной литературе автоматизированные системы тестирования знаний называют *CAT*-системами (*Computer Adaptive Testing Systems*).

Вообще говоря, в настоящее время известны компьютерные дидактические программы следующих типов:

- обучающие программы;

- тестирующие (контролирующие) программы;

- моделирующие программы, требующие от обучающегося воспроизведения последовательности рассуждений или «сборки» правильного результата на основе знаний, предоставленных системой;

- программные тренажеры, предназначенные для отработки и закрепления технических навыков решения задач;

- дидактические игры, предполагающие выдачу ответов обучающимся на формируемые системой вопросы в игровой форме;

- гипертекстовые системы – мультимедийные справочники, обладающие развитой системой навигации и поиска информации.

Современные развитые АОС включают в себя все или большую часть указанных программ. Как минимум, современные АОС решают следующие педагогические задачи:

а) демонстрация учебного материала в различных формах;

б) тренинг в изучаемой области, позволяющий закрепить полученный материал;

в) тестирование и диагностика, позволяющие оценить степень усвоения обучающимся учебного материала, а также осуществить контроль всего процесса обучения.

Подчеркнем принципиальную недостаточность использования компьютерных технологий только для проверки знаний обучающегося. Правильнее говорить об *обучающем тестировании*, которое предполагает тесную интеграцию процесса обучения и контроля обучения [5].

С применением методов искусственного интеллекта на основе обучающего тестирования создаются адаптивные АОС, реализующие личностно-ориентированный подход к процессу обучения. Данный подход означает индивидуализацию обучения, когда каждый из обучающихся имеет возможность выбора оптимального для него темпа, времени и, вообще, условий обучения, возможность выбора уровня трудности тестовых заданий и т.д. [6]. В ходе обработки результатов обучающего тестирования преподаватель может получить оперативную информацию не только об особенностях выполнения определенных заданий обучающимся, но и об его индивидуальных характеристиках - показателях умственного развития, скоростных особенностях, динамики работоспособности и т.д.

В адаптивные АОС процесс обучения рассматривается, как процесс функционирования системы автоматизированного управления [6]. Обратная связь в интеллектуальных адаптивных АОС осуществляется с помощью непосредственных реплик (указаний или запросов) обучающегося. Однако основной контур обратной связи в таких системах реализуют на основе постоянного контроля уровня его знаний.

Критерии оценки качества тестов рассматриваются в работе с точки зрения перспективности применения их в рамках обучающе-тестирующей подсистемы АОС.

В первом разделе работы рассматривается общие вопросы, связанные с тестовым методом контроля качества обучения. Второй раздел содержит обзор различных способов классификации тестов и тестовых заданий. В разделах с третий по шестой последовательно рассматриваются основные критерии оценки надежности, валидности, трудности и дискриминативности тестов. В заключении сформулированы основные выводы.

Для простоты записи вместо слов «обучающийся», «обучаемый», «испытуемый» и т.д. будем далее писать «ученик», а вместо слов «преподаватель», «эксперт», «обучающий» и т.д. – «учитель».

1. Тестовый метод контроля качества обучения

1.1. Теоретические предпосылки. Классическая тестология базируется на следующих теоретических посылках [6].

1) Способности человека врожденны и в силу этого фактически неизменны. Эта предпосылка является центральной, именно она приводит к сопоставлению способностей, со знаниями, умениями и навыками.

2) Высокий уровень способностей встречается редко, причем способности у людей распределяются в соответствии с законом Гаусса.

3) Игнорирование качественного развития человека (как умственного, так и психического), сведение различия в способностях разных людей к чисто количественным показателям.

4) Принцип "решил – не решил", т. е. контролю подлежит лишь конечный результат деятельности, а особенности интеллектуальной деятельности при выполнении задания не диагностируются и, соответственно, не учитываются.

В настоящее время указанные подходы к тестированию пересматриваются. Все большее значение придается роли обучения в развитии способностей. Происходит отказ от чисто количественного подхода к возрастному развитию интеллекта. Способности человека рассматриваются, как продукт прижизненного формирования. Определяющая роль в этом процессе отводится обучению. В этих условиях главной функцией контроля качества обучения становится определение условий, наиболее благоприятствующих дальнейшему развитию данного человека. Этот подход также снимает противопоставление способностей знаниям, умениям и навыкам. Знания – это всегда элемент какой-то деятельности, а умения, навыки, способности – это всегда деятельность (действие и система действий), характеризующаяся определенными способностями.

1.2 Принципы тестирования. Вслед за работой [3] выделим следующие основные принципы тестирования:

- связь с целями обучения;
- объективность;
- справедливость и гласность;
- систематичность;
- гуманность и этичность;
- научность и эффективность.

Принцип связи с целями обучения. Из целей обучения следует, что цели тестирования, а, говоря более широко, и педагогических измерений вообще, должны отвечать критериям социальной полезности и значимости, научной корректности и общественной поддержки.

Принцип объективности. Использование в педагогических измерениях этого принципа призвано уменьшить влияние субъективизма и предвзятости в процессе этих измерений.

Принцип справедливости и гласности педагогических измерений означает одинаково доброжелательное отношение ко всем ученикам, открытость всех этапов процесса измерений, своевременность ознакомления учеников с результатами измерений.

Принцип систематичности предполагает систематичность тестирований и самопроверок каждого учебного модуля, раздела и каждой темы. Важным аспектом данного принципа является требование репрезентативного представления содержания учебного курса в содержании теста.

Принцип гуманности и этичности педагогических измерений означает, что тестовые задания и процедура тестирования должны исключать нанесение какого-либо вреда ученикам, не допускать ущемления их по национальному, этническому, материальному, расовому, территориальному, культурному и другим признакам. Этот же принцип означает, что тестирование может быть только добровольным.

Принцип научности и эффективности предписывает необходимость проверки содержания и правильности формы тестов независимыми экспертами-учителями по соответствующим учебным дисциплинам.

Важнейшим является принцип, в соответствии с которым тесты должны быть построены по методике, обеспечивающей выполнение требований соответствующего государственного образовательного стандарта.

К принципам тестирования примыкают принципы построения тестовых заданий, включающие в себя следующие принципы [6].

Коллегиальная подготовка тестовых заданий. Коллегиальное построение, оценивание и выбор тестовых заданий независимыми учителями позволяет существенно уменьшить важнейший недостаток индивидуального контроля знаний – его субъективность. Отметим, что с формальной точки зрения в этой связи возникает проблема многокритериальной оценки и выбора заданий группой лиц, принимающих решения.

Централизованное накопление тестовых заданий. Составленные и отобранные экспертами тестовые задания должны храниться в базе данных системы тестирования, обрабатываться учителями по соответствующей дисциплине с целью устранения возможных дублирований заданий.

Унификация инструментальных средств подготовки тестовых заданий. Образовательные учреждения должны использовать унифицированное программное обеспечение систем тестирования, инвариантное к предметной области.

1.3 Методические аспекты контроля знаний связаны с решением педагогических и психологических вопросов [2]. К методическим аспектам контроля знаний относятся:

− выбор типов и трудности тестовых заданий («что контролировать?»);

− планирование процедуры контроля знаний («когда контролировать?»);

− формирование набора адекватных тестовых заданий («как контролировать?»).

*Выбор типов и трудности тестовых заданий***.** Набор тестовых заданий должен соответствовать цели контроля на данном этапе учебного процесса. Так на этапе восприятия, осмысления и запоминания оценивается уровень знаний ученика о предметной области и понимания основных положений. Способность ученика применять полученные знания для решения конкретных задач, требующих проявления познавательной самостоятельности, оценивается как соответствие требуемым навыкам и/или умениям.

*Планирование процедуры контроля знаний***.** Учебный процесс принято рассматривать как распределенный во времени процесс формирования требуемых знаний, навыков и умений. Соответственно этому, выделяют следующие четыре этапа контроля знаний.

1) Исходный (предварительный) контроль. Данный контроль проводится непосредственно перед обучением, имея целью оценить начальный уровень знаний ученика и соответственно планировать его обучение.

2) Текущий контроль. Осуществляется в ходе обучения и позволяет определить уровень усвоения учеником отдельных разделов учебного материала, а затем на этой основе скорректировать дальнейшее изучение предмета.

3) Рубежный контроль. Проводится по завершении определенного этапа обучения и служит цели оценки уровня знаний ученика по теме или разделу курса.

4) Итоговый контроль. Позволяет оценить знания, умения и навыки ученика по курсу в целом.

Формирование набора адекватных тестовых заданий. Важным методическим аспектом контроля знаний является формирование набора тестовых заданий, в зависимости от вида и цели контроля. Различные методы формирования заданий для контроля (случайная последовательность заданий разной сложности, последовательность заданий в порядке нарастания их уровня сложности и т.д.) рассмотрены ниже.

1.4 Проблемы тестового метода. Вслед за работой [3] назовем основные проблемы педагогических измерений, вообще, и тестового метода контроля знаний, в частности:

- латентность измеряемых свойств учеников;
- сложность выбора числа и состава индикаторов качества обучения;
- неоднозначность концептуализация измеряемых свойств учеников;
- трудность операционализации свойств учеников.

Латентность измеряемых свойств учеников означает, что эти свойства доступны учителю лишь в неявной форме и недоступны для непосредственного измерения. Примерами латентных свойств ученика являются «качество подготовленности ученика», «уровень знания учеником данной учебной дисциплины», «уровень интеллектуального развития ученика» и т.п.

Выбор числа и состава индикаторов. Вследствие латентности измеряемых свойств учеников эти свойства приходится измерять косвенно, через эмпирически фиксируемые проявления некоторых признаков (индикаторов) знаний. Обычно каждое из заданий теста рассматривают как индикатор, предназначенный для выявления какого-либо одного аспекта знаний учеников.

Неоднозначность концептуализации. Опять же в силу латентности измеряемых свойств учеников в педагогических измерениях одной из первых возникает задача концептуализации этих свойств. Традиционно, концептуализация измеряемых свойств учеников осуществляется в терминах знаний, умений, навыков и представлений. В компетентностном подходе к образованию используется интегральная концепция «компетентность». Важной частью процесса концептуализации является определение возможных источников погрешностей измерения рассматриваемых свойств ученика.

Трудность операционализации. Операционализация некоторого свойства ученика выражается в правилах измерения этого свойства, таких, например, как «должен знать принципы…», «должен знать методы…», «должен знать формулы….», «должен уметь применять формулы….» и т.д. Можно сказать, что операционализация формирует прагматическое определение знания учебного курса вида «ученик удовлетворительно знает данный учебный курс, если он правильно отвечает на такие-то тестовые задания».

В связи с проблемой совершенствования тестового метода контроля качества обучения актуальным является решение следующих задач [7].

а) Разработка методик коллегиального экспертного построения тестовых заданий, а также оценивания степени истинности вариантов ответов на них относительно многозначной лингвистической шкалы.

б) Разработка моделей и методик оценивания степени согласованности коллегиального мнения учителей о характеристиках созданного с их участием набора тестовых заданий.

в) Разработка моделей и методик количественного оценивания объективности процесса педагогического контроля.

г) Разработка моделей нечеткого и статистического оценивания знаний, позволяющие организовать эффективное педагогическое тестирование по гуманитарным, общественно-политическим и другим дисциплинам, характеризующимся существенной диалектичностью знаний.

В работе [6] выделена еще одна проблема в области тестирования, требующая дальнейших исследований – проблема определения оптимального времени выполнения заданий. Ограниченность этого времени в большинстве современных тестов создает

неравные условия для учеников, поскольку результаты тестирования отражают не только уровень знаний и умений учеников, но и различную скорость их нервных процессов. В результате в проигрыше оказываются следующие две группы учеников: лица с хорошо отработанными умениями, необходимыми для решения заданий теста, но с малым индивидуальным темпом; ученики, возможно, обладающие высоким природным темпом, но не имеющие хорошо отработанных навыков и умений. В этом плане особенно актуальной является разработка всех аспектов адаптивного тестирования, которое позволяет проводить испытания практически в индивидуальном темпе.

К проблемам тестового метода следует отнести также проблему идентификации личности ученика [8]. Пусть мы уверены, что ученик не мошенничает во время проверки знаний, но как узнать, что с тестовой системой взаимодействует именно данный ученик? Для идентификации личности ученика достаточно визуального контакта, но существуют и другие способы.

Отметим, наконец, проблему распознавание ответов ученика [6]. Самой сложной при этом является задача диагностирования характера и природы его ошибок. В общем случае такое "распознавание" ответа требует знания, как внутридисциплинарных, так и междисциплинарных связей. В целом, анализ семантики ответов ученика представляет собой одну из наиболее важных и сложных задач математического обеспечения интеллектуальных адаптивных АОС.

1.5. Методы тестового контроля можно разделить на

- неадаптивные методы,
- частично адаптивные методы,
- полностью адаптивные методы [2].

Неадаптивные методы. Общим для всех неадаптивных методов является то, что в процессе контроля все ученики проходят одну и ту же последовательность проверочных заданий. Эта последовательность не зависит от действий ученика во время контроля, так что всем им выдаются задания одинаковой трудности либо в виде фиксированного набора, либо случайным образом. Число заданий является постоянным для всех учеников, не зависимо от их уровня подготовленности.

Частично адаптивные методы тестового контроля знаний предполагают, что последовательность и число тестовых заданий различно для сильных, средних и слабых учеников. Выбор тестирующей системой числа и трудности заданий происходит с учетом ответов ученика и/или на основе подготовленного учителем сценария проведения контроля знаний, использующего математическую модель ученика.

Адаптивные методы максимально используют ту или иную математическую модель ученика, а также модель учебного материала (в форме, например, семантической сети соответствующей онтологии). Данные методы позволяют организовать индивидуальный контроль знаний каждого ученика, поддерживая оптимальный для него уровень трудности выдаваемых контрольных заданий, формируя индивидуальные стратегии контроля по отдельной теме, разделу или курсу в целом и т.д.

1.6. Методы оценки знаний. Проверка знаний учеников может быть осуществлена на основе различных критериев формирования оценки. В зависимости от этого методы оценки знаний можно разделить на

- методы на основе количественных критериев,
- методы на основе вероятностных критериев,
- методы на основе классификационных критериев [7].

Методы на основе количественных критериев предполагают использование количественной шкалы, т.е. оценка в этом случае задается числом. В простейшем случае эта оценка может представлять собой сумму баллов, полученных учеником за правильные ответы на тестовые задания. В более сложных случаях при формировании оценки учитывают типы и характеристики тестовых заданий.

В *методах на основе вероятностных критериев* главным является определение вероятности правильного ответа ученика, как функции уровня его подготовленности и параметров тестового задания (раздел 6.2).

Методы на основе классификационных критериев предполагают отнесение ученика к одному из нескольких устойчивых классов с учетом совокупности признаков, определяющих данного ученика. Примерами методов этого класса являются методы на основе алгоритма вычисления оценок (ABO), а также методы на основе нечетких множеств.

1.7. Шкалы оценок. Одной из проблем, сопровождающих разработку тестов, является проблема выбора шкалы для оценки качества выполнения тестовых заданий. Традиционно знания учеников оцениваются с использованием лингвистической шкалы оценок. Очевидно, что при формировании такой шкалы велика доля субъективизма, обусловленного различиями в опыте, интуиции, компетентности, профессионализме и уровне требовательности учителя.

Первичной информацией при тестировании знаний является набранный балл ученика (первичный балл). Достоинством этой оценки является ее простота и наглядность. Однако, легко видеть, что первичный балл является не абсолютной, а относительной оценкой. Так эта оценка существенно зависит, к примеру, от трудности заданий теста. Таким образом,

желательна объективная шкала оценок уровня подготовленности учеников, подтверждаемая на различных тестах, имеющих заранее определенный уровень трудности заданий. Вторым существенным недостатком первичных баллов является их нелинейность по отношению к уровню подготовленности учеников.

Проблематика построения шкал оценок, свободных от указанных и других недостатков, детально рассмотрена, например, в работ [4].

1.8. Этапы разработки тестов. При разработке тестов следует соблюдать ряд правил, обеспечивающих создание надежного и сбалансированного инструмента оценки знаний.

В первую очередь, необходимо проанализировать содержание тестовых заданий с позиции равной представленности в тесте разных фрагментов учебного курса. Тест не должен быть нагружен второстепенными терминами и несущественными деталями. Задания теста должны быть сформулированы четко, кратко и недвусмысленно, чтобы все ученики могли понять смысл того, что от них требуется. Ни одно из заданий теста не должно служить подсказкой для ответа на другое задание.

Варианты ответов на каждое задание должны быть подобраны таким образом, чтобы исключались возможности простой догадки или отбрасывания заведомо неподходящего ответа. Варианты ответов также следует формулировать кратко и однозначно. Удобна альтернативная форма ответов, когда ученик должен выбрать одно из перечисленных решений “да-нет”, “верно-неверно”.

Тестовые задачи не должны быть слишком громоздкими или слишком простыми. Вариантов ответов на задачу должно быть, по возможности, не менее пяти. В качестве неверных ответов желательно использовать наиболее типичные ошибки учеников.

Естественной является следующая последовательность действий при разработке теста:

- определить цель тестирования, отобрать материал для теста;
- выбрать подходы к процессу разработки, создать план теста и его спецификацию;
- разработать тестовые задания и выполнить их экспертный анализ;
- провести пробное тестирование и проанализировать его результаты;
- выбрать критерии оценки качества теста;
- в соответствии с выбранными критериями произвести оценку качества теста;
- выполнить доработку теста и его параллельных форм.

2. Классификация тестов

2.1. Подходы к разработке тестов. Отметим прежде, что различают два основных подхода к разработке тестов:

- нормативно-ориентированный подход;

- критериально – ориентированный подход.

Нормативно-ориентированный подход позволяет выполнить сравнение учебных достижений (уровней знаний, умений и навыков) учеников друг с другом на основе распределения баллов.

Критериально-ориентированный подход ориентирован на оценку того, в какой степени ученик овладел необходимым для профессиональной деятельности учебным материалом.

Нормативно-ориентированные и критериально-ориентированные тесты отличаются, в первую очередь, целями создания тестов. Если первые из них позволяют оценить соответствие знаний, умений и навыков ученика некоторой норме – «подходит - не подходит», то вторые - ориентированы на оценку уровня обученности ученика и эффективности программы обучения. Второе различие между указанными тестами состоит в уровне детализации предметной области. От критериально–ориентированных тестов чаще всего требуется большая детализация. Третье различие заключается в процедуре обработки результатов тестирования. Итоговые баллы по результатам нормативно–ориентированного тестирования базируются на статистических данных нормативной группы учеников и формируются с использованием специальных нормативных шкал.

2.2. Классы тестов. Хотя процессы усвоения теории и овладения навыками по решению задач взаимосвязаны и взаимообусловлены, тем не менее, на практике они протекают достаточно не зависимо друг от друга. Основываясь на этом, в работе [9] предложена следующая иерархическая классификация тестов:

- тесты распознавания;

- ученического применения;

- механического воспроизведения;

- алгоритмического применения;

- фрагментарного понимания;

- эвристического применения;

- целостного понимания;

- творческой деятельности.

Тесты распознания предполагают контроль возможности ученика выделить (распознать) необходимую информацию из предложенного набора. Тесты данного класса

ориентированы на выявление правильности самых общих представлений ученика об изучаемом предмете.

Тесты ученического применения. Ученику в этом случае предлагается решение задач, в которых заданы цель, ситуация и его возможные действия. От ученика требуется, например, выбрать формулы для нахождения требуемых величин. Подчеркнем, что тесты этого класса предполагают наличие внешней подсказки.

Механическое воспроизведение. Ученик по памяти должен воспроизвести требуемое знание, не объясняя его. Например, ученик должен механически воспроизвести определение, формулировку закона и т.п. Внешняя подсказка при этом отсутствует. Материал в тесте данного класса воспроизводится в такой форме, в которой он дан учителем или содержится в учебнике.

Алгоритмическое применение предполагает контроль возможности ученика использовать абстракции (правила, законы, методы и т.п.) для решения типовых задач изучаемого предмета.

Тесты фрагментарного понимания исходят из того, что ученик должен правильно понимать правила, законы, методы и т.п., но лишь в пределах конкретного раздела учебного курса (без необходимости увязки его с материалом других разделов курса).

Тесты эвристического применения ориентированы на проверку умения ученика использовать абстракции (правила, законы, методы и т.п.) для решения эвристических задач, в которых задана цель, но не определены пути ее достижения. От ученика в этом случае требуется умение идентифицировать ситуацию и применить необходимый ранее усвоенный материал.

Целостное понимание предполагает проверку возможности ученика в полной мере использует такие логические операции как синтез и анализ. При подготовке ответа на тесты данного класса ученику необходимо использовать знания, относящиеся ко всей изучаемой дисциплине или ее разделу.

Тесты творческой деятельности призваны дать возможность ученику продемонстрировать умение решать задачи творческого типа, добывать объективно новую информацию.

2.3. Виды тестов. Выделяют следующие виды тестов [1]:

- вербальные и невербальные тесты;
- групповые и индивидуальные тесты;
- тесты достижений;
- личностные тесты;

- объективные тесты;

- проективные тесты;

- простые тесты;

- сложные тесты;

- интегративные тесты.

Тест на словарный запас представляет собой пример вербального теста. Невербальный тест требует определенных действий в качестве ответа.

Тест достижений. Примерами тестов достижений являются тесты успеваемости, тесты творческих возможностей, тесты способностей, сенсорно-моторные тесты, тесты интеллекта.

Личностные тесты представляют собой тесты на установки, интересы, темперамент, характерологические и мотивационные тесты.

Объективные тесты, в отличие от субъективных, проводятся в условиях, когда ученик знает о действительной цели исследования.

Проективные тесты предполагают ответ, который не может быть расценен как "правильный" или "не правильный". Эти тесты используют такой способ построения тестового задания, при котором ученик не должен выбирать ответ из заданного списка.

Простой (гомогенный) тест предназначен для оценки качества знаний по одной учебной дисциплине.

Сложные (гетерогенные) тесты, в отличие от простых тестов, состоят из нескольких самостоятельных подтестов, по каждому из которых должен быть получен ответ ученика и на этой основе сформирована общая оценка. Таким образом, гетерогенный тест включает в себя несколько гомогенных тестов (иногда говорят «шкал») и охватывает содержание нескольких дисциплин.

Интегративный тест состоит из заданий, ответы на каждое из которых требует знания нескольких учебных дисциплин.

2.4. Формы тестовых заданий. Вслед за работой [10] выделим следующие формы тестовых заданий:

- цепные задания;

- тематические задания;

- текстовые задания;

- ситуационные задания.

Цепные задания. Цепными называются задания, в которых правильный ответ на последующее задание зависит от ответа на предыдущее задание.

Тематические задания представляют собой совокупность тестовых заданий любой формы, разработанных для контроля знаний учеников по одной изученной теме. Задания могут быть цепными и тематическими одновременно, если их цепные свойства имеют место в рамках одной темы.

Текстовые задания – это совокупность заданий, созданных для контроля знаний учениками конкретного учебного текста. Текстовые задания широко применяются при изучении русского и иностранных языков. Например, с их помощью можно организовать проверку знания стихотворений, грамматики, синтаксиса и др. В виде текстовых заданий можно также оформить тест на усвоения понятийного состава учебной дисциплины. Кроме того, текстовые задания удобны для проверки классификационных знаний.

Ситуационные задания в работе [10] определены, как педагогически переработанный фрагмент профессиональной деятельности специалиста Ситуационные задания разрабатываются для проверки знаний и умений учеников действовать в практических, экстремальных и других ситуациях. Эти задания также хорошо подходят для интегрального контроля уровня знаний учеников.

Каждая из рассмотренных форм тестовых заданий имеет несколько вариантов. Например, возможны задания с выбором одного правильного ответа, с выбором одного наиболее правильного ответа и задания с выбором нескольких правильных ответов. Отметим, что последний вариант является наиболее предпочтительным.

2.5. Типы вопросов. Можно выделить четыре типа вопросов, используемых в тестовых заданиях [10]:

- закрытая форма;
- открытая форма;
- установление соответствия;
- установление последовательности.

Закрытая форма является наиболее распространенной и предлагает несколько альтернативных ответов на поставленный вопрос. Например, ученику задается вопрос, требующий альтернативного ответа «да» или «нет», «является» или «не является», «относится» или «не относится» и т. п. Тестовое задание, содержащее вопрос в закрытой форме, включает в себя один или несколько правильных ответов и иногда называется выборочным заданием.

Закрытую форму вопросов используют также в тестах-задачах с выборочными ответами. В тестовом задании в этом случае формулируют условие задачи и все необходимые исходные данные, а в ответах представляют несколько вариантов результата

решения в числовом или буквенном виде. Ученик должен решить задачу и показать, какой из представленных ответов он получил.

*Открытая фор*ма. Вопрос в открытой форме представляет собой утверждение, которое необходимо дополнить. Данная форма может быть представлена в тестовом задании, например, в виде словесного текста, формулы (уравнения), чертежа (схемы), графика, в которых пропущены существенные составляющие - части слова или буквы, условные обозначения, линии или изображения элементов схемы и, графика. Ученик должен по памяти вставить соответствующие элементы в указанные места («пропуски»).

Установление соответствия. В данном случае ученику предлагают два списка, между элементами которых следует установить соответствие.

Установление последовательности предполагает необходимость установить правильную последовательность предлагаемого списка слов или фраз.

В работе [11] предложена несколько иная классификация типов вопросов в тестовых заданиях:

- выборочный ответ;
- числовой ответ;
- проверка простой формулы;
- проверка логической формулы;
- проверка слова.

Выборочный ответ. Вопрос данного типа формулируется так, что на него можно привести набор вариантов ответов, каждый из которых некоторым образом кодируется (цифрой, символом, набором символов, картинкой и т.п.). Среди предлагаемых вариантов ответов может быть один или несколько правильных. Ученик должен указать либо все верные ответы, либо их нужное число.

Числовой ответ. Ученик в данном случае должен решить задачу или произвести некоторые действия, в результате которых должно получиться число.

Проверка простой формулы. Ответ ученика в этом случае имеет вид не очень сложной формулы, правильность которой можно проверить простым способом, например сопоставлением результатов вычислений по введенной учеником и правильной формулам.

Проверка логической формулы. Здесь в ответ на поставленный вопрос ученик должен ввести некоторую последовательность слов или выражений. Правильность ответа в этом случае тестирующая программа проверяет с помощью некоторой логической формулы.

Проверка слова предполагает ввод учеником последовательности слов или других символов в ответ на открытый вопрос, содержащий пропуски этих слов или символов.

3. Критерии надежности тестов

Надежностью теста называется степень совпадения его результатов при повторном тестировании одних и тех же учеников в одинаковых или близких условиях. Известно значительное число критериев надежности теста [12]. Например, в качестве такого критерия можно использовать коэффициент корреляции Пирсона между двумя параллельными тестами на одной и той же выборке учеников. Однако повторная проверка знаний по одному и тому же вопросу связана с лишней психологической нагрузкой учеников и их переутомлением. Кроме того, создание истинно параллельных тестов практически нереально.

В качестве критерия надежности можно использовать также коэффициент корреляции результатов тестирования и результатов экспертных опенок. К лишней психологической нагрузке учеников и их переутомлению в этом случае добавляется необходимость организации группы экспертов и, тем самым, увеличению нагрузки на учителей.

Чаще всего в качестве критерия надежности тестового задания используют коэффициент надежности Гутмана и коэффициент корреляции Спирмана-Брауна, а также их некоторые модификации.

Далее нам понадобятся следующие обозначения:

$P = (p_1, p_2, ..., p_N) = (p_i, i \in [1:N])$ - множество учеников, где i - номер ученика, N – их общее число;

$T = (t_1, t_2, ..., t_M) = (t_j, j \in [1:M])$ - набор тестовых заданий теста T, где j - номер тестового задания в тесте, M – общее число заданий;

$X = (x_{i,j}, i \in [1:N], j \in [1:M])$ - матрица результатов тестирования, где $x_{i,j}$ - оценка i-го ученика за выполнение j-го задания.

Критерии надежности теста обычно строятся на основе следующих величин:

$y_i = \sum_{j=1}^{M} x_{i,j}$ - суммарный тестовый балл ученика p_i по результатам выполнения теста T;

$\tilde{r}_i = \frac{r_i}{M}$, $\tilde{w}_i = \frac{w_i}{M}$ - доли правильных и неправильных ответов, где r_i, w_i - суммарные числа правильных и неправильных ответов, полученных учеником p_i по результатам выполнения теста T соответственно;

$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$ - средний арифметический балл по всем ученикам (т.е. оценка математического ожидания случайной величины y_i);

$D = \frac{s^2}{N-1}$ - оценки дисперсии тестовых результатов всех учеников, где $s^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2$ - сумма квадратов отклонений баллов учеников от своих средних значений;

$\sigma = \sqrt{D}$ - соответствующая оценка средних квадратичных отклонений.

3.1. Коэффициент надежности Гутмана. Одними из самых простых методов оценки надежности тестов являются методы, основанные на оценке структурированности знаний ученика. Наиболее известным методом этого класса является метод Гутмана (*L.L. Guttman*).

Положим, что задания в тесте T расположены в порядке возрастания их сложности и матрица X является бинарной, т.е. оценки $x_{i,j}$ могут принимать только значения ноль и единица. Строка с номером i этой матрица образует, так называемый, профиль ученика p_i, характеризующий структуру его знаний. При идеальной структуре теста (когда сложность заданий действительно возрастает с увеличением их номера) и идеальной структуре знаний ученика p_i «правильным» будет профиль, в котором сначала идут только единицы, а затем – только нули. Отклонение оценки $x_{i,j}$ от правильного профиля назовем ошибкой профиля и обозначим $e_{i,j}$. Точнее говоря, положим, что $e_{i,j} = 0$, если оценка $x_{i,j}$ является «правильной», и $e_{i,j} = 1$ - в противном случае.

В случае, когда структура теста является идеальной, любая ошибка профиля означает либо незнание ученика, либо его неудачную попытку угадать правильный ответ. При усреднении результатов тестирования по всем ученикам последний эффект неизбежно нивелируется и существенными оказываются лишь инверсии (нарушения верной последовательности) в заданиях теста. На этом основании Гутманом была в качестве критерия надежности теста предложена величина

$$\gamma = \gamma(P,T) = \frac{\sum_{i,j} e_{i,j}}{NM}, \qquad (1)$$

называемая коэффициентом надежности Гутмана. Здесь суммирование ведется по всем $i \in [1:N]$, $j \in [1:M]$ и сумма имеет смысл общего числа ошибочных ответов всех учеников. Иногда величину (1) называют коэффициентом структурированности тестовых результатов. В качестве нижней допустимой границы коэффициента надежности Гутмана обычно принимают величину 0,8.

На основе профилей учеников предложены и другие критерии качества теста, например, коэффициент правильности профиля [12].

3.2. Коэффициент корреляции Спирмана-Брауна. Метод оценки надежности тестов с помощью коэффициента корреляции Спирмана-Брауна (*Spearman-Brown*) основан на идее

оценки стабильности результатов учеников. Метод относится к классу методов раздельного коррелирования и является наиболее часто используемым методом этого класса [3].

Введем еще следующие обозначения:

y_i^e, y_i^o - суммарные тестовые баллы ученика p_i по результатам выполнения нечетных и четных заданий теста T соответственно, $y_i^e + y_i^o = y_i$;

$\bar{y}_e = \frac{\sum_{i=1}^{N} y_i^e}{N}$, $\bar{y}_o = \frac{\sum_{i=1}^{N} y_i^o}{N}$ - средние арифметические баллы по всем ученикам (т.е. оценки математических ожиданий случайных величин y_i^e, y_i^o соответственно);

$D_e = \frac{s_e^2}{N-1}$, $D_o = \frac{s_o^2}{N-1}$ - оценки дисперсий тестовых результатов y_i^e, y_i^o соответственно, где $s_e^2 = \sum_{i=1}^{N} (y_i^e - \bar{y}_e)^2$, $s_o^2 = \sum_{i=1}^{N} (y_i^o - \bar{y}_o)^2$ - суммы квадратов отклонений баллов ученика p_i от своих средних значений;

$\sigma_e = \sqrt{D_e}$, $\sigma_o = \sqrt{D_o}$ - оценки средних квадратичных отклонений величин y_i^e, y_i^o;

$K_{eo} = \frac{s_{eo}}{N}$ - оценка корреляционного момента (момента связи) величин y_i^e, y_i^o, где $s_{eo} = \sum_{i=1}^{N} (y_i^e - \bar{y}_e)(y_i^o - \bar{y}_o)$ - сумм произведений отклонений величин y_i^e, y_i^o от своих средних значений;

$r_{eo} = \frac{K_{eo}}{\sigma_e \sigma_o} \approx \frac{s_{eo}}{\sqrt{s_e^2 s_o^2}}$ - оценка коэффициента корреляции величин y_i^e, y_i^o;

$\bar{\varepsilon} = \frac{\sum_{i=1}^{N} \varepsilon_i}{N}$ - среднее значение ошибки ε_i (т.е. оценки математического ожидания случайной величины ε_i), где $\varepsilon_i = y_i^e - y_i^o$ - ошибка ученика p_i по результатам выполнения нечетных и четных заданий теста T;

$D_\varepsilon = \frac{\sum_{i=1}^{N} (\varepsilon_i - \bar{\varepsilon})^2}{N-1}$ - оценки дисперсии ошибок ε_i.

Коэффициентом корреляции Спирмана-Брауна называется величина

$$\eta_{eo} = \frac{2r_{eo}}{1 + r_{eo}}. \qquad (2)$$

Полагается, что тест достаточно надежен при $\eta_{eo} > 0{,}8$.

Вариантом формулы (2) является формула

$$\widetilde{\eta}_{eo} = 1 - \frac{D_\varepsilon}{D},$$

где, напомним, D_ε, D - оценки дисперсии ошибки тестирования и дисперсии тестовых результатов всех учеников соответственно.

Некоторые другие варианты формулы (2) рассмотрены, например, в работе [12]. В этой же работе рассмотрены методы оценки надежности тестов на основе использования результатов дисперсионного и факторного анализа. Для гомогенных тестов, например, с помощью дисперсионного анализа можно получить *индекс надежности теста* и *индекс гомогенности теста*, как критерии его надежности.

Рассмотренный коэффициент корреляции Спирмана-Брауна получен с помощью расщепления теста (*split-half method*). Другими, гораздо менее удобными и реже используемым является метод двух параллельных тестов (*parallel-form reliability*) и метод повторного тестирования с помощью одного и того же теста (*test-retest reliability*) [4].

3.3. Коэффициент надежности *KR*-20. Широкое распространение на практике получило применение для расчета надежности теста так называемой формулы *KR*-20

$$K = \frac{M}{M-1}\left(1 - \frac{\sum \widetilde{r}_i \widetilde{w}_j}{D}\right),$$

где суммирование ведется по всем $i \in [1:N]$, $j \in [1:M]$. Формула получила свое название по имени ее создателей *F. Kuder* и *M. Richardson* [13] (число 20 означает номер формулы в указанной публикации).

Полагается, что если величина коэффициента надежности *K* составляет от 0,90 до 0,99, то тест имеет отличную оценку надежности, если от 0,80 до 0,89 - то хорошую, от 0,70 до 0,79 – удовлетворительную, менее 0,69 - неудовлетворительную надежность. Для текущего контроля знаний рекомендуется использовать тест, имеющий коэффициента надежности не менее 0,80, а для итоговой аттестации - более 0,90 [4]. Отметим, что по формуле *KR*-20 оценивается надежность таких широко известных тестов как *SAT* и *TOEFL*.

4. Критерии валидности теста

Валидность (от англ. *validate*) характеризует способность теста давать результаты, позволяющие осуществить их правильную интерпретацию с точки зрения цели тестирования [14]. В АОС обычно используют различные подходы на основе эмпирического внутреннего метода оценки валидности теста [15]. Рассмотрим подходы этого класса, построенные на основе следующих характеристик теста [12]:

- нормальность распределения результатов тестирования;

- ошибки регрессионной модели;

- плотность покрытия тестом учебного материала.

4.1. Валидность по распределению. При данном подходе тест считается валидным, если средний результат тестирования присущ, большей части учеников, а сами результаты распределяются по нормальному закону [12]. Очевидно, что в соответствии с данным критерием валидный тест должен содержать подавляющую долю заданий средней трудности, но он обязательно должен включать в себя и явно легкие и явно трудные задания. Валидность теста по распределению достигается путем замены тестовых заданий, которые нарушают нормальность распределения результатов тестирования. Оценка отклонения результатов тестирования от нормального закона распределения представляет собой классическую задачу теории вероятностей. Для решения этой задачи могут быть использованы все методы, предлагаемые этой теорией.

4.2. Валидность на основе ошибок регрессионной модели. Идея этого подхода состоит в следующем [2].

а) На основе тестирования первой части рассматриваемой группы учеников строим уравнение регрессии вида $\tilde{x} = f(j)$, где j - номер тестового задания, $\tilde{x}$ - прогнозируемая оценка за тестовое задание t_j. При этом в качестве искомой функции $f(j)$ обычно используется полиномиальная функция вида

$$f(j) = a_m^* j^m + a_{m-1}^* j^{m-1} + \ldots + a_1^* j + a_0^*$$

с неизвестными коэффициентами $A^* = (a_m^*, a_{m-1}^*, \ldots, a_1^*, a_0^*)$, которые определяются из условия

$$\min_{A \in R^m} \left(x - \tilde{x}(A)\right)^2 = \left(x - \tilde{x}(A^*)\right)^2,$$

где $A = (a_m, a_{m-1}, \ldots, a_1, a_0)$, R^m - m-мерное арифметическое пространство.

б) На основе тестирования второй части группы учеников (естественно, с использованием того же набора заданий) проверяем адекватность полученного уравнения регрессии $\tilde{x} = f(j)$. Если прогнозируемые оценки близки к фактическим, делаем вывод о валидности рассматриваемого набора тестовых заданий. В качестве критерия валидности в этом случае можно использовать, например, максимальное отклонение предсказанной оценки от фактической.

4.3. Валидность по плотности покрытия (в работе [12] - валидность по содержанию). Поскольку целью теста является оценка знаний учениками того или иного учебного материала, тестовые задания, в идеале, должны охватывать все разделы и темы этого учебного материала, все понятия и связи между понятиями и т.д.

Положим, что базе знаний АОС предметная область рассматриваемой учебной дисциплины представлена в виде простой (не расширенной) семантической сети [16]. О плотности покрытия можно говорить на уровне модуля, библиотеки моделей и учебного курса. Рассмотрим в качестве примера тестирование знаний на уровне модуля *m*, семантическая сеть которого определена ориентированным графом без контуров *G*. Вершины этого графа соответствуют входным понятиям модуля $\bar{\mathbf{C}}$. и его выходным понятиям $\mathbf{C} = \{c_i, i \in [1, n]\}$, где $n \geq 0$ - общее число выходных понятий С. Дуги графа *G* соответствуют отношению «определяемое понятие – определяющее понятие», которое связывает понятия наборов $\bar{\mathbf{C}}$, С в семантическую сеть. Таким образом, идеальный тест для модуля *m* должен включать в себя проверку знаний испытуемыми всех *n* выходных понятий модуля С, а также знаний всех связей этих понятий с другими понятиями наборов $\bar{\mathbf{C}}$, С.

Для оценки валидности по плотности теста *T*, предназначенного для проверки знаний испытуемыми модуля *m*, можно предложить следующие критерии.

а) Критерий $\alpha_1 = \frac{\tilde{n}}{n}$ - относительное число выходных понятий модуля *m*, знания которых проверяет тест *T*. Здесь $\tilde{n}$ - абсолютное число таких понятий.

б) Критерий $\alpha_2 = \sum_{i=1}^{n} \alpha_{2,i}$ - относительное число отношений, связывающих выходных понятий модуля *m* с другими понятиями этого модуля, знания которых проверяет тест *T*. Здесь $\alpha_{2,i} = \frac{\tilde{k}_i}{k_i}$ - проверяемое тестом *T* относительное число отношений, связывающих выходное понятие $c_i \in \mathbf{C}$ с теми понятиями этого модуля, с которыми данное понятие информационно связано в узком смысле; $\tilde{k}_i$ - абсолютное число таких понятий, k_i - общее число отношений, информационно связывающих понятие c_i с другими понятиями этого модуля в узком смысле [17].

в) Аналогичный критерию α_2 критерий $\alpha_3 = \sum_{i=1}^{n} \alpha_{3,i}$, в которой рассматриваются понятия, информационно связные с данным понятием в широком смысле [17].

г) Критерий $\alpha_4 = \rho_1\alpha_1 + \rho_2\alpha_2 + \rho_3\alpha_3$ - аддитивная свертка критериев $\alpha_1, \alpha_2, \alpha_3$; ρ_1, ρ_2, ρ_3 - весовые коэффициенты.

Может оказаться целесообразной некоторая нормализация рассмотренных критериев. Очевидны также более «тонкие» варианты этих критериев, учитывающие сложность

соответствующих концептов и отношений [17]. Наряду с критерием α_4, можно предложить множество аналогичных критериев, построенных на основе других сверток [16, 17].

Несколько в другой форме аналогичные критерии валидности теста можно построить на основе, так называемой, *матрица контрольных заданий* [18]. Рассмотрим, как и ранее, учебный модуль *m*. Обозначим *k* общее число отношений $D = \{d_i,\ i \in [1:k]\}$ –, связывающих в узком смысле все выходные понятия этого модуля с другими понятиями $\bar{\mathbf{C}}$, С. Матрица контрольных заданий представляет собой булеву $(M \times (n+k))$-матрицу *Q*. Здесь, напомним, что *n* – число выходных концептов рассматриваемого модуля, *M* – число строк матрицы, которые соответствуют тестовым заданиям. Единица в позициях $j_1, j_2, \ldots, j_p$ строки *i* матрицы *Q* означает, что тестовые задания, соответствующие этой строке, позволяют проверить правильность усвоения учеником выходных концептов c_{j_1}, $c_{j_2}, \ldots, c_{j_p}$. Аналогично ноль в оставшихся позициях строки не позволяет проверить правильность усвоения учеником оставшихся концептов. Здесь полагается, что $j_l \in [1:n]$, $l \in [1:p]$; $p \in [1:M]$. Аналогично, единица в позициях $j_1, j_2, \ldots, j_q$ строки *i* матрицы *Q* означает, что соответствующие тестовые задания позволяют проверить правильность усвоения учеником связей c_{j_1}, $c_{j_2}, \ldots, c_{j_q}$. Здесь полагается, что $j_l \in [(n+1):)n+k)]$, $l \in [1:q]$; $q \in [1:M]$.

5. Критерии трудности теста

Обычно в качестве критерия трудности теста используют *индекс трудности теста* λ, который определяется относительным числом учеников, давших правильный ответ на данной тест $\tilde{r}$:

$$\lambda = (1 - \tilde{r}). \qquad (3)$$

В психологии различают субъективную и объективную трудности теста. Субъективная трудность определятся индивидуальными психологическими и иными характеристиками тестируемого, например, лимитом времени, доступностью инструкции, его психическим состоянием. Формула (3) определяет объективную трудность теста. Если, например, на данный тест в среднем правильно отвечает 20% учеников ($\lambda = 0{,}8$), то данный тест следует классифицировать, как трудный. Если, напротив, верный ответ дают 80% учеников ($\lambda = 0{,}2$), тест классифицируется, как легкий.

Отметим, что, не смотря на свое название, объективная трудность теста зависит от особенностей выборки учеников (возрастные, профессиональные, социокультурные различия и т.д.). Отметим также, что индекс трудности теста применим лишь к заданиям, для

которых можно определить «правильный» и «неправильный» ответы. Например, при определении с помощью тестов личностных характеристик ученика, понятие трудности заданий теста, как правило, неприменимо.

Для успеха тестирования с помощью теста очень важным является подбор заданий по индексу трудности. При выборе слишком трудных заданий резко снижаются валидность и надежность теста. С другой стороны, слишком простые задания приводят к незначительной вариативности их результатов, что делает затруднительным корректное оценивание учеников.

Обычно задания с низким индексом трудности помещают в начале теста, а задания с высоким индексом – ближе к концу теста. Несколько самых легких заданий размещают перед основными заданиями теста и используют в качестве примера. В то же время, в тестах скорости часто используют задания с относительно невысокими и примерно одинаковыми индексами трудности. При этом общее число заданий выбирается таким, чтобы никто из учеников за заданное время не успел решить все их.

6. Критерии дискриминативность теста

Дискриминативность (разрешающая способность) теста характеризует его способность отделить испытуемых с высокой продуктивностью учебной деятельности от испытуемых с низкой продуктивностью.

6.1. Коэффициент дискриминации. Простейшим критерием дискриминативности теста является *коэффициент дискриминации* r_d, определяемы по следующей схеме.

а) Проводят тестирование достаточно большой группы учеников ($N > 100$).

б) На основе результатов тестирования отбирают примерно по 27% учеников в лучшую группу P_{best} и худшую группу P_{worst}, содержащие по N_b и N_w учеников соответственно.

в) Вычисляют числа правильных ответов v_b, v_w в указанных группах соответственно.

г) Вычисляют коэффициент дискриминации

$$\varpi_d = \frac{v_b}{N_b} - \frac{v_w}{N_w},$$

где $\frac{v_b}{N_b}, \frac{v_w}{N_w}$ - доли правильных ответов в группах P_{best}, P_{worst}.

д) Если $r_d > 0{,}3$, то полагают, что дискриминативность рассматриваемого теста является достаточной.

Критерии дискриминативности теста можно построить также на основе матрица корреляций тестовых заданий с тестовыми баллами испытуемых [12].

6.2. Разрешающая способность теста. Весьма содержательные критерии дискриминативности теста можно построить на основе модели Г. Раша(*G. Rasch*)

$$p_{i,j} = \frac{\exp\left(\delta(\theta_i - \beta_j)\right)}{1 + \exp\left(\delta(\theta_i - \beta_j)\right)}, \qquad (4)$$

где $p_{i,j}$ - вероятность правильного ответа *i*-го ученика на тестовое задание t_j; θ_i - латентный уровень знаний этого ученика; β_j - латентный уровень трудности тестового задания t_j; δ - нормирующий множитель; $i \in [1:N]$, $j \in [1:M]$ [3]. Часто используется множитель $\delta = 1{,}7$, обеспечивающий совместимость модели Раша (4) с известной моделью *A. Fergusson* [3].

Величины θ_i, β_j измеряют в логитах l_{θ_i}, l_{β_i}, где

$$l_{\theta_i} = \ln\frac{\widetilde{r}_i}{\widetilde{w}_i}, \quad l_{\beta_i} = \ln\frac{\widetilde{R}_i}{\widetilde{W}_i}.$$

Здесь

$$\widetilde{r}_i = \frac{r_i}{M}, \ \widetilde{w}_i = \frac{w_i}{M} \ -$$

доли правильных и неправильных ответов, полученных *i*-м учеником по результатам выполнения теста *T* соответственно, $i \in [1:N]$;

$$\widetilde{R}_j = \frac{R_j}{N}, \ \widetilde{W}_j = \frac{W_j}{N},$$

где R_j, W_j - соответственно, суммы правильных и неправильных ответов всех учеников при выполнении тестового задания t_j; $j \in [1:M]$.

Итерационная процедура оценки значений величин l_{θ_i}, l_{β_i} (параметрическая идентификация модели) детально рассмотрена, например, в работах [3, 12]. Данная процедура включает в себя вычисление оценок средних значений уровня подготовленности учеников и трудности заданий, а также дисперсий и стандартных отклонений этих оценок. Процедура обычно строится на основе метода наибольшего правдоподобия Р. Фишера. Однако могут использоваться и другие, более эффективные методы нахождения устойчивых оценок значений указанных латентных параметров.

Разрешающую способность теста можно оценить длиной промежутка Δθ, измеренной в логитах, на латентной шкале уровня подготовленности, который соответствует разности первичных баллов, равных единице. Неравенство $|\theta_1 - \theta_2| \le \Delta\theta$ означает, что данный тест не в состоянии различить уровни знаний ученика, равные θ_1 и θ_2.

Для оценки разрешающей способности теста можно использовать также, так называемый, *точечно-бисериальный коэффициент корреляции* ϖ_{bis}, выражающий связь

между результатами ответов ученика на данное задание теста с индивидуальными баллами выборки учеников [3].

Заключение

В работе приведены основные сведения о тестовом методе контроля качества обучения – теоретические предпосылки метода, принципы тестирования, методические аспекты тестирования и т.д. Значительное внимание уделено классификации тестов. Основное содержание работы составляет обзор критериев оценки таких аспектов качества теста, как надежность, различительная способность (валидность), вариативность, трудность, дискриминативность. Как отмечалось выше, работа носит преимущественно обзорный характер, однако в работе также предложен ряд критериев валидности теста по плотности покрытия, основанных на модели знаний предметной области изучаемой дисциплины в виде семантической сети.

Результаты работы показывают, что имеется значительное число критериев для оценки каждого из указанных аспектов качества тестов и тестовых заданий. В то же время, в подавляющем большинстве современных АОС производится не многокритериальная, а однокритериальная оценка уровня знаний учеников. Таким образом, актуальной является задача разработки методов, алгоритмов и соответствующих программ, реализующих в АОС многокритериальную оценку и ранжирование учеников. Некоторые аспекты указанной задачи рассмотрены в работе [19].

Интенсивно развиваемый в настоящее время компетентносный подход к образованию ставит на повестку дня вопрос о разработке методов контроля компетентности учеников. Некоторые аспекты этих методов рассмотрены, например, в работе [8]. Современные АОС включают в себя все больше элементов, вплоть до элементов виртуальной реальности, поддерживающих деятельностное обучение [20]. Данное направление не рассматривается в работе. В работе на рассмотрена также проблематика контроля знаний в случае использования рейтинговой системы оценки знаний [21]. Отметим, наконец, заслуживающую самостоятельного исследования проблему контроля понятийных знаний ученика. Некоторые подходы к решению этой проблемы рассмотрены, например, в работах [22 - 26].

Авторы выражают благодарность Добрякову А.А. за ценные рекомендации и советы, имеющие отношение к предмету данного обзора.

Работа выполнена в рамках Государственного контракта №16.740.11.0407.

Литература

1. Истоки экспериментальной психологии (http://www.effecton.ru/199.html).

2. Прокофьева Н.О. Вопросы организации компьютерного контроля знаний // (Международный электронный журнал). Educational Technology & Society 9(1) 2006, pp.433 – 440. (http://www.ifets.info/index.php?http://www.ifets.info/main.php).

3. Аванесов В.С. Тесты в социологическом исследовании/ В.С. Аванесов.- М.: Изд-во «Наука», 1982. – 199с.

4. Ким В.С. Тестирование учебных достижений. Монография. - Уссурийск: Издательство УГПИ, 2007. - 214 с.

5. Углев, В. А. Обучающее компьютерное тестирование // Теоретические и прикладные вопросы современных информационных технологий: Материалы VIII Всероссийской научно-технической конференции. - Улан-Удэ: ВСГТУ, 2007. - С. 312 – 316.

6. Латышев В.Л. Интеллектуальные обучающие системы: контроль знаний и психодиагностика. (http://nit.miem.edu.ru/2004/plenar/9.htm).

7. Рудинский И.Д., Аскеров Э.М., Емелин М.А., Строилов Н.А. Принципы и технологии создания интегрированной автоматизированной системы контроля знаний // Информационные технологии в образовании и науке: Сб. трудов ВНПК. - М., 2006, С. 17-35.

8. Мякишев В.В., Семченко В.И. Контроль знаний в системе дистанционного обучения. (http://www.1th.ru/ru/tools/articles/contrznanii/).

9. Екимова Л.С., Мосин Ю.В., Рогожина Т.С., Ромашин С.Н., Тарасова М.А. Разработка электронной системы обучающих тестов / Тезисы докладов всероссийской научно-практической конференции «Информационные технологии в образовании и науке "ИТОН-2006", Москва, 4-5 февраля 2006 года. (www.iton.mfua.ru/2006/tesis/all.html).

10. Аванесов В.С. Форма тестовых заданий. Учебное пособие. Второе издание -М.: Центр Тестирования, 2005. –155с.

11. Кривицкий Б.Х. К вопросу о компьютерных программах учебного контроля знаний // (Международный электронный журнал). Educational Technology & Society 7(2) 2004, pp.158 – 169. (http://ifets.ieee.org/russian/periodical/V_72_2004EE.html).

12. Олейник Н.М. Тест как инструмент измерения уровня знаний и трудности заданий в современной технологии обучения. Учебное пособие.: Донецк, Донецкий Государственный Университет. (http://opentest.com.ua/test-kak-instrument-izmereniya-urovnya-znanij/).

13. Kuder G.F., Richardson, M.W. The theory of the estimation of test reliability // Psychometrika, 1937, v.2, N3. -p.151-160.

14. Общая психология. Словарь.
(http://slovari.yandex.ru/ ~книги/Общая психология/Валидность теста/).

15. Валидность (http://ru.wikipedia.org/wiki/%C2%E0%EB%E8%E4%ED%EE%F1%F2%FC).

16. Карпенко А.П., Соколов Н.К. Расширенная семантическая сеть обучающей системы и оценка ее сложности // Наука и образование: электронное научно- техническое издание, 2008, 12. (http://technomag.edu.ru/doc/111716.html).

17. Карпенко А.П., Соколов Н.К. Оценка сложности семантической сети в обучающей системе // Наука и образование: электронное научно- техническое издание, 2008, 11. (http://technomag.edu.ru/doc/106658.html).

18. Обучение и контроль с использованием ИТ. (http://www.bestreferat.ru/referat-53297.html).

19. Аскеров Эмин Мубариз оглы. Автоматизация многокритериального оценивания уровня сформированности профессиональных компетенций будущих специалистов. Автореферат дис. ... кандидата технических наук : 05.13.06 [Место защиты: Ин-т информатизации образования] Москва, 2010, 19 с.

20. Кофтан Ю.Р. Контроль в компьютерных обучающих средах / Материалы XV Международной конференции "Применение новых информационных технологий в образовании", 29-30 июня 2004г.г. Троицк, Московской области. (http://ict.edu.ru/vconf/files/7289.doc).

21. Хасанова Е.В. Методика формирования рейтинговой системы оценки знаний для повышения качества обучения школьников с использованием сетевых методов обработки информации. (http://www.tgc.ru/conf/region/?2000_2/02.html).

22. Пустобаев В.П., Саяпин М.Ю. Фомализация элементов диагностики знаний учащихся // Информатика и образоввани, 2005, №2, с. 120 -123.

23. Галямова Е.В., Карпенко А.П., Соколов Н.К, Ягудаев Г.Г. Контроль понятийных знаний субъекта обучения в обучающей системе // Вестник МАДИ (ГТУ). -М., 2009. – 2(17). – С.82-86.

24. Гагарина Д.А., Калмыков А.А. Семантическая родовидовая сеть понятий и экспертная система для ее тестирования. (http://tm.ifmo.ru/tm2004/db/doc/get_thes.php?id=218).

25. Гагарина Д.А. Информационная система для организации самостоятельного изучения понятий и контроля тестовых знаний (http://kafedratsp.narod.ru/nd_nauka/konf_2006/s2_3.doc).

26. Калмыков А.А. Экспертная система для оценивания понятийного состава знаний обучаемых. (http://www.nsu.ru/archive/conf/nit/97/c1/node5.html).