

Муханова А.А., Муханов С.А.

**ТЕХНОЛОГИЯ ПРОЕКТИРОВАНИЯ И ОЦЕНКИ ТЕСТОВ В
СДО MOODLE**

*Московский государственный машиностроительный университет (МАМИ),
107023, г. Москва, ул. Б.Семёновская, д. 38.*

Mukhanova A.A., Mukhanov S.A.

**TECHNOLOGY OF DESIGN AND ASSESSMENT OF TESTS IN
LMS MOODLE**

*Moscow state university of mechanical engineering (MAMI),
107023, Moscow, Bolshaya Semenovskaya, 38.*

В данной статье рассматривается технология проектирования критериально-ориентированных тестов, а также оценка тестов с использованием встроенных механизмов СДО Moodle и их интерпретация в теории педагогических измерений.

Ключевые слова: педагогический тест, система дистанционного обучения, Moodle, теория педагогических измерений.

This article discusses the technology of designing criterion-oriented tests and evaluation tests using the built-in mechanisms LMS Moodle and their interpretation in the theory of educational measurement.

Keywords: pedagogical test, distance learning system, Moodle, educational measurement theory.

Впервые метод тестирования был применен в психологии в конце XIX века. Родоначальником данного метода считают Джеймса Маккина Кэттела. Основы психологического тестирования были описаны Фрэнсисом Гальтоном в

книге «Исследование человеческих способностей и их развитие». Алфред Бине известен как автор первого практического теста интеллекта.

Впервые появившись в психологии данный метод был взят на вооружении и педагогами. Так, например, Уильямом Штерном разрабатывалась концепция интеллектуального коэффициента, которая позднее легла в основу известного теста IQ Альфреда Бине. Таким образом, первые работы по теории тестов появились на стыке психологии, социологии, педагогики и других поведенческих наук (Behavioral Sciences).

В СССР психологические и педагогические тесты получили распространение в 60-х годах XX века в работах С.И. Воскерчьян, Э.И. Моносзон, В.П. Беспалько и др.

В энциклопедиях данному термину дается такое определение: «тест» (от англ. слова *test* — «испытание», «проверка») — метод изучения глубинных процессов деятельности человека, посредством его высказываний или оценок факторов функционирования системы управления.

В контексте данной статьи нас будут интересовать, прежде всего, педагогические тесты. В отечественной литературе имеется различные определения педагогических тестов, в частности С.И. Воскерчьян считает, что «тест — это кратковременное, технически просто обставленное испытание, проводимое в равных для всех условиях и имеющее вид такого задания, решение которого поддается количественному учету и служит показателем степени развития к данному моменту известной функции у данного испытуемого». Наиболее развернутое определение, на наш взгляд дают И.А. Рапопорт и др.: «тест — это самым тщательным образом подготовленная в соответствии с определенными разработанными правилами, прошедшая предварительную экспериментальную проверку и специальную процедуру для ее улучшения, имеющая достаточные характеристики своей эффективности совокупность вопросов и заданий, предъявляемых испытуемому с целью квалитметрического выявления социальных, психических и психофизиологических характеристик его личности, отличающаяся

формализацией ответов испытуемых, выделением в них части, несущей наибольшую информационную нагрузку, что ускоряет, облегчает и объективизирует их последующий анализ, обработку и интерпретацию». [1]

В современной России обращает на себя внимание поляризация мнений: от горячего одобрения тестов, до их обостренного неприятия. Между тем, тестирование уже прочно вошло в современную школу на совершенно разных уровнях: от начальной и средней школы до использования тестов в высшей школе. Наиболее яркими примерами, конечно, служат тесты ГИА и ЕГЭ в средней школе и тесты ФЭПО в высшей школе. Не ставя своей целью выявлять сильные и слабые стороны тестирования, а также не вступая в дискуссию по их целесообразности отметим, что проверка знаний в данной форме широко используется во всем мире. Одним из наиболее известных тестов, наверное, является тест GMAT (англ. Graduate Management Admission Test) — стандартизованный тест для определения способности успешно обучаться в бизнес-школах. GMAT используется наиболее уважаемыми школами бизнеса по всему миру как один из критериев отбора, чаще всего для приема на программу MBA.

Существует несколько подходов к проектированию тестов, связанных с проектированием всего учебного курса. Наиболее часто используется метод проектирования «сверху - вниз». Суть данного метода заключается в том, что первоначально строится генеральное содержание учебного предмета с разбивкой на укрупненные модули (дидактические единицы). Затем проводится детализация модулей на более простые образования. Крайне желательно, чтобы указанные образования уже в своей формулировке были диагностируемы. Как нам кажется, для этих целей отлично подходит понятия «микроцели», сформулированное в педагогической технологии академика В.М. Монахова, а также требования, предъявляемые к ней. Микроцель – главные вопросы темы, зоны ближайшего развития учащихся. Микроцели должны быть конкретны, сформулированы в такой форме, которая может быть диагностируема,

например, в форме: «знать», «уметь», «понимать», «иметь представление о», «уметь давать характеристику». Одна микроцель рассчитана на 4-6 часов.

Каждая микроцель связана с остальными микроцелями курса и все вместе они составляют семантический граф. Наиболее простой способ проектирования теста заключается в составлении вопросов к понятиям, составляющим узлы семантического графа, а также разработка практических упражнений, основанных на свойствах данных понятий. Усложнением здесь является разработка заданий, определяющих отношения между понятиями или дидактическими единицами в целом.

Что касается проектирования самих заданий, то, по мнению В.С. Аванесова [2], к заданиям в тестовой форме предъявляются следующие требования:

- правильность содержания;
- логическая форма высказывания;
- правильность формы;
- краткость;
- наличие определенного места для ответов;
- правильность расположения элементов задания;
- одинаковость правил оценки ответов;
- одинаковость инструкции для всех испытуемых;
- адекватность инструкции форме и содержанию задания.

Проектирование теста следует начать с определения сферы его применения. По данному параметру тесты распадаются на два больших класса - критериально-ориентированные (criterion-referenced test) и нормативно-ориентированные (norm-referenced test) [3].

Критериальный тест позволяет выявить степень усвоения испытуемым определённого раздела курса. Может случиться, что все учащиеся выполнили тест (это означает, что они успешно освоили учебный материал), или что все не выполнили тест – материал не освоен. Целью нормативно-ориентированного теста является упорядочение испытуемых по уровню их подготовленности. При

этом сами баллы, полученные учащимися при выполнении теста не имеют значения для целей теста, т.к. основная цель – ранжирование участников. Возможны случаи, когда какое-то задание не дифференцирует испытуемых, например, задание легкое – и все успешно его выполнили. И наоборот, очень трудное задание – и все его не выполнили. Такие задания не позволяют провести ранжирование и поэтому должны быть удалены из теста. Мы будем рассматривать, в основном, критериально-ориентированные тесты.

Помимо собственно проектирования содержания теста необходимо также уделить значительное внимание получению объективных характеристик, позволяющих оценить качество тестов и их составляющих – вопросов теста. Одно из несомненных достоинств тестов и, прежде всего, в компьютерном варианте в том, что их можно использовать неоднократно. Это позволяет при оценке качества теста использовать статистические методы при вычислении ряда характеристик, использующихся как для их оценке в целом, так и для оценки отдельных вопросов теста. В основе системы сбора статистики используется метрическая система Раша (Rash Measurement), позволяющая говорить о качестве педагогических измерений.[4]

Для анализа тестовых результатов в рамках Rash Measurement необходимы исходные данные в виде матрицы, которая обычно является бинарной, т.е. состоящей из столбцов, соответствующих тестовым заданиям, и строк, соответствующих испытуемым. На пересечении строк и столбцов находится единица, если испытуемый справился с данным заданием, и ноль в противном случае.

В качестве инструмента для проектирования тестов и организации системы тестирования нами была выбрана СДО (Система дистанционного обучения) Moodle (аббревиатура от «модульная объектно-ориентированная динамическая учебная среда»). Данная СДО комбинирует в себе несколько классов систем:

- Система управления сайтом (CMS).
- Система управления обучением (LMS).
- Виртуальная среда обучения (VLE).

Moodle предлагает широкий спектр возможностей для полноценной поддержки процесса обучения в дистанционной среде - разнообразные способы представления учебного материала, проверки знаний и контроля успеваемости.

В качестве инструмента для организации тестирования Moodle предлагает следующие возможности:

- разные форматы вопросов: множественный выбор (единственный или множественный варианты правильного ответа), альтернативный вопрос (верно/неверно), вопрос на соответствие, числовой и вычисляемый вопрос и др.;
- использование в вопросах картинок и иных мультимедийных объектов, использование формул в формате TeX;
- перемешивание вариантов ответа в случайном порядке, отбор случайных вопросов из базы заданий теста;
- задание ранга ответам, что позволяет задать более сложным заданиям более высокий уровень оценки;
- выделение групп вопросов и возможность отбора в итоговый тест определенного количества вопросов из группы, что позволяет в итоге предъявить учащемуся определенное количество вопросов, нацеленных на проверку определенного аспекта (микроцели или дидактической единицы, в зависимости от целей теста). Также данный подход позволяет сгруппировать вопросы по уровню сложности.

Еще одной замечательной возможностью, предоставляемой Moodle является то, что по результатам теста система автоматически генерирует весьма содержательный отчет, который позволяет произвести анализ статистических показателей, полученных по тесту, в том числе и с использованием Rash Measurement, вернее, система выдает уже рассчитанные показатели, полученные при помощи данной системы.

Чтобы посмотреть результаты тестирования и рассчитанные параметры теста преподаватель должен открыть в системе сам тест и затем перейти на

вкладку «Результаты». Помимо самих оценок, выставленных в границах, заданных преподавателем для теста при щелчке на пункт «Анализ вопросов» откроется окно со статистическими параметрами теста.

Основными характеристиками теста в целом, которые рассчитываются по известным из статистики формулам, являются (в скобках указано название параметра в системе Moodle в случае, если оно не переведено):

Средняя оценка испытуемых, представляющая собой среднее арифметическое по оценкам всех выполняющих тест студентов.

Медиана (Median grade) - срединное значение оценок испытуемых.

Стандартное отклонение оценок за тест (Standard deviation) – это мера вариации полученных при прохождении теста баллов. Данная величина характеризует пригодность теста для дифференциации учащихся по уровню усвоения материала. Рассчитывается она по формуле (1):

$$S_x = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1}} \quad (1)$$

Коэффициент асимметрии (Score distribution skewness) показывает смещение максимума влево (отрицательная асимметрия) или вправо (положительная) от положения максимума на нормальной кривой.

Вообще, считается, что хороший нормативно-ориентированный тест обеспечивает нормальное распределение индивидуальных баллов репрезентативной выборки учащихся, когда среднее значение баллов (примерно около 70%) находится в центре распределения, а остальные значения концентрируются вокруг среднего по нормальному закону. Таким образом, положительное значение асимметрии указывает на то, что, возможно, тест излишне легкий, отрицательное – излишне сложный.

Эксцесс (Score distribution kurtosis) – мера крутизны («островерхости») кривой. При положительном значении эксцесса кривая имеет более острую вершину, чем кривая нормального распределения, при отрицательном значении – наоборот. Применительно к тестам следует обратить внимание на то, что

положительное значение эксцесса говорит о том, что у теста не очень велика дифференцирующая способность. Еще одним аспектом является то, что понятие эксцесса имеет смысл в том случае, если у распределение только одна мода. Полимодальное распределение будет говорить о том, учащиеся, сдававшие тест распределились на несколько групп.

Коэффициент внутренней согласованности теста (Coefficient of internal consistency) - одна из разновидностей коэффициента надежности теста, вычисляемая по формуле Кьюдера-Ричардсона (2) как коэффициент корреляции между результатами ответов испытуемых на четные и нечетные вопросы одного и того же теста.

$$R_n = \frac{N}{N-1} \frac{\sum_{i=1}^N x_i^2}{\sum_{i=1}^N x_i} - \frac{\sum_{i=1}^N x_i^2}{\sum_{i=1}^N x_i} \quad (2)$$

где N – число задач в тесте; s_x^2 – дисперсия первичных оценок теста; p – индекс трудности заданий; $g = 1-p$; (p и g – представляют долю испытуемых, соответственно справившихся и не справившихся с заданием). Такая надежность вычисляется по данным однократного применения теста. Коэффициент надежности можно интерпретировать в процентах дисперсии показателей определяемой разными источниками. Так, коэффициент надежности равный 0,7 показывает, что 70% дисперсии результатов теста зависит от истинной дисперсии по измеряемому свойству, а 30% - от дисперсии ошибки. Дисперсия ошибки включает неоднородность заданий теста, временные показатели, изменение состояния испытуемых, влияние тренировки и другие факторы.

Стандартная ошибка (Standard error) — оценивает фактор везения и указывает границы погрешности для оценки студента за тест. Так, например, если стандартная ошибка 12% и студент выполнил 70% заданий теста, тогда его подлинная оценка будет располагаться в интервале от 58% до 82%.

Считается, что в хорошем тесте средний арифметический балл равен медианному значению оценок, кривая распределения оценок имеет нормальный вид, а значения среднего арифметического, моды и медианы совпадают. Это

признак хорошо подобранного уровня трудности теста, адекватного подготовке учащихся.

Еще одна группа параметров позволяет провести оценивание конкретных тестовых заданий:

Индекс легкости (Facility index) - процент студентов, которые ответили на конкретный вопрос теста правильно.

Стандартное отклонение (Standard deviation) - характеризует разброс значений оценок испытуемых при ответе на конкретный вопрос теста.

Случайно угаданная оценка (Random guess score) - оценка, которую может получить испытуемый случайным угадыванием.

Предполагаемый вес (Intended weight) - вес, назначенный данному вопросу преподавателем. Здесь вес вопроса рассчитывается не в баллах, как это делается на этапе редактирования теста, а в процентах.

Эффективный вес (Effective weight) - характеристика того, какова фактическая доля данного вопроса в итоговой оценке. Данная характеристика позволяет скорректировать преподавателю вес вопроса.

Коэффициент дискриминации (Discrimination index) - это соотношение связи между оценками за конкретный вопрос и за весь тест в целом. Т.е., для хорошего тестового вопроса, студенты с наивысшими оценками за него так же будут иметь более высокие оценки и за тест в целом.

Для расчета коэффициента дискриминации, по результатам ответов на все вопросы теста, студенты делятся на три части: сильные, средние и слабые. Математическое выражение для индекса дискриминации i -го вопроса имеет вид:

$$I_d = \frac{(X_{i\phi} - X_{i\psi})}{N} \quad (3)$$

где $X_{i\phi}$ - сумма дробей баллов, полученных при ответе на i -й вопрос одной третью испытуемых, которые получили самые высокие баллы за тест в целом; и $X_{i\psi}$ - сумма относительных значений баллов, полученных одной третью самых слабых испытуемых по результатам всего теста; N - общее число испытуемых.

Данный коэффициент может приобретать значение в диапазоне от -100% до 100%. Например, 100% означает, что на данный вопрос все сильные студенты дали правильный, а все слабые - неправильный ответ. Значение коэффициента около 0 означает, что сильные и слабые студенты отвечали на данный вопрос одинаково. Если значение коэффициента отрицательно, то скорее всего вопрос содержит ошибку, так как он показывает, что слабые студенты дали ответы лучше, чем сильные.

Данный коэффициент является важной статистической характеристикой дифференцирующей способности тестовых заданий, которую позволяют вычислять средства Moodle. Считается, что задание обладает достаточной дифференцирующей способностью, если значение данного коэффициента имеет значение больше или равное 30%.

Эффективность дискриминации (Discriminative efficiency) - нормированный коэффициент дифференциации. В этом коэффициенте тоже сопоставляются ответы сильных и слабых студентов. Формула для расчета:

$$K_d = \frac{\sum a_{ix}}{NS_i S_x} \quad (4)$$

где $\sum a_{ix}$ - сумма произведений отклонений набранных баллов для данного вопроса и теста в целом,

N - количество ответов на данный вопрос,

S_i - стандартное отклонение набранных баллов в этом вопросе,

S_x - стандартное отклонение набранных баллов для теста в целом.

Данный коэффициент принимает значения, аналогичные предыдущему, но дает более точные результаты, чем коэффициент дискриминации, так как при вычислении коэффициента принимают во внимание результаты всей группы студентов, не вдаваясь искусственному делению на трети сильных, средних и слабых.

Статистическая оценка базы тестовых вопросов, позволяет корректировать тест.

Таким образом, рекомендации по проектированию тестов могут быть следующие:

1. Провести разбиение курса на модули (дидактические единицы).
2. Каждую дидактическую единицу представить в виде ее логической структуры на основе проектирования учебных микроцелей.
3. Разработать систему тестовых заданий по каждой микроцели. Тестовые задания разрабатываются с учетом целей теста: критериально-ориентированный или нормативно-ориентированный.
4. Провести первичную экспертизу тестовых заданий.
5. Организовать тестирование с целью собрать статистику по тесту, затем провести анализ расчетных показателей, характеризующих как тест в целом, так и отдельные его вопросы. Выполнить корректировку теста и отдельных заданий.

База тестовых заданий должна быть достаточно велика, т.к. малое количество заданий приводит к ненадежной оценке качества теста и создает условия для невозможности применения методик его улучшения. Основное качество теста – его валидность. Существуют четыре основных способа установления валидности теста: очевидная, содержательная, конструктивная и прогностическая валидность. Для целей данной статьи нас преимущественно интересует содержательная валидность. Оценка содержательной валидности предполагает, что тест обнаруживает в себе полный набор заданий для измерения определенной черты. Малое количество вопросов, очевидно, не будет отвечать данному параметру.

При проектировании итогового теста по курсу рекомендуется разделить вопросы на группы по дидактическим единицам, а также, при необходимости, по конкретным микроцелям курса. СДО Moodle позволяет построить тест таким образом, что в конкретный вариант теста, предъявляемый учащемуся, будет отобрано необходимое, установленное преподавателем, количество вопросов из каждой группы.

Рассмотренная система проектирования тестов может служить инструментом повышения эффективности разработки контролируемых материалов как на этапе их разработки, так и в дальнейшем при их оценке в учебных заведениях.

Литература:

1. Рапопорт И.А., Сельг Р., Соттер И. Тесты в обучении иностранным языкам в средней школе. – Таллин: Валгус, 1987. – 350 с.

2. Аванесов В.С. Теория и практика педагогических измерений (материалы публикаций)

3. Ким В.С. Тестирование учебных достижений. Монография. - Уссурийск: Изд.УГПИ, 2007. - 214 с. – [Электронный ресурс]. Режим доступа: http://www.uspi.ru/static/kim_testing_monograph/ [Дата обращения: 21 сентября 2013].

4. Аванесов В.С. Применение тестовых форм в Rasch Measurement. – Педагогические измерения, 2005, N 4. - С. 3-20.

5. Коржик И.А., Протасова И.В., Толстобров А.П. Некоторые методы увеличения качества сценария теста в Moodle // Сборник докладов Международной интернет-конференции «Информационно-технологическое обеспечение образовательного процесса государств-участников СНГ» [Электронный ресурс]. Режим доступа: <http://elib.bsu.by/handle/123456789/27794> [Дата обращения: 15 сентября 2013].

References:

1. I.A. Rapoport, Selg R., J. Sotter Tests in teaching foreign languages in high school. - Tallinn: Valgus, 1987. - 350 p.

2. Avanesov V.S. Theory and practice of educational measurement (materials publications)

3. Kim V.S. Testing of academic achievement. Monograph. - Ussurijsk: UGPI, 2007. - 214 p. - [Electronic resource]. Mode of access:

http://www.uspi.ru/static/kim_testing_monograph/ [Access Date: 21 September 2013].

4. Avanesov V.S. The use of test forms to Rasch Measurement. - Pedagogical measurement 2005, N 4. - P. 3-20.

5. Korzhik I.A., Protasova I.V., Tolstobrov A.P. Some methods of increasing the quality of the test script in Moodle // Proceedings of the International Internet Conference "Information Technology of the educational process of the CIS" [electronic resource]. Mode of access: <http://elib.bsu.by/handle/123456789/27794> [Access Date: 15 September 2013].

Дата отправки: 21.09.2013

© Муханова А.А., Муханов С.А.