УДК 378.147.88

МЕТОДИКА АНАЛИЗА И ПОВЫШЕНИЯ КАЧЕСТВА ТЕСТОВ В СИСТЕМЕ ЭЛЕКТРОННОГО ОБУЧЕНИЯ MOODLE

И. В. Протасова, А. П. Толстобров, И. А. Коржик

Воронежский государственный университет

Поступила в редакцию 18.07.2014 г.

Аннотация. Разработана и апробирована методика оценки качества тестовых контрольно-измерительных материалов в системе электронного образования Moodle, основанная на подходах современной теории педагогических измерений. Выявлены факторы, влияющие на надежность тестового задания, выработаны рекомендации по улучшению качества тестов.

Ключевые слова: электронный образовательный ресурс, Moodle, тестирование, статистика, Item Response Theory.

Annotation. The technique of an assessment of quality of test control and measuring materials in system of electronic formation of Moodle based on approaches of the modern theory of pedagogical measurements is developed and approved. The factors influencing reliability of a test task are revealed, recommendations about improvement of quality of tests are developed.

Keywords: e-learning systems, Moodle, testing, statistics, Item Response Theory.

ВВЕДЕНИЕ

Внедрение компьютерных технологий в образование послужило толчком к созданию автоматизированных средств оценки уровня успешности освоения учебного материала с помощью тестов, которые являются одной из наиболее технологичных форм проведения такого контроля с управляемыми параметрами качества. Формализованная тестовая система оценка знаний основывается не на единичном задании, субъективно оцениваемом преподавателем, а на широком спектре вопросов-заданий, охватывающих различные аспекты изучаемого учебного материала. Статистическая обработка результатов таких испытаний дает возможность получения числовых характеристик, позволяющих объективно оценивать как результаты испытуемых, так и качество тестовых заданий и теста в целом. Принципы такой обработки результатов тестовых испытаний определяются различными моделями теории педагогических измерений [1].

Современная, широко используемая в мировом образовательном сообществе система электронного обучения Moodle [2], также эффективно используемая в Воронежском государственном университете, содержит развитую тестовую подсистему с встроенным механизмом автоматизированной статистической обработки результатов тестирования и вычисления показателей качества тестовых материалов. Однако, к сожалению, возможно в силу сложности интерпретации и недостатка информации о ее реализуемости, использование результатов такой обработки в реальной практике электронного обучения остается весьма ограниченным.

Целью данной работы является проведение анализа реальных результатов тестовых испытаний по двум дисциплинам образовательных программ факультета компьютерных наук ВГУ, основанного на моделях современной теории педагогических измерений, для сравнения возможностей практического применения этих моделей и выработки рекомендаций и создания методики по проведению анализа и интерпретации показателей качества тестовых контрольно-измеритель-

ных материалов в практике использования конкретной системы электронного обучения.

МОДЕЛИ СТАТИЧЕСКОЙ ТЕОРИИ ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ

Статистические теории педагогических измерений имеют более чем вековую историю развития. Разработанная в 1904 году Чарльзом Спирманом статистическая модель оценки истинного балла испытуемых явилась основой целостной «классической» статистической теории измерений [3]. В 1963 году под руководством Ли Кронбаха создана расширенная статистическая теория Generalizability Theory (GT-теория) [3]. Как классическая, так и расширенная статистические теории имеют общий предмет рассмотрения – оценку истинных и ошибочных компонент педагогического измерения с целью определения надежности тестовых результатов. Обе эти теории применимы к обработке данных на выборках небольшого объёма. Однако у них есть и важные различия [1–5]. Расширенная статистическая теория измерений позволяет вести поиск нескольких источников систематических и случайных погрешностей измерения [1–5]. Если в классической теории предполагается, что дисперсии истинных и ошибочных компонентов измерения, а также корреляции с внешним критерием в параллельных вариантах тестирования являются одинаковыми, то в GT–теории, параллельные варианты теста считаются случайными выборками из одной генеральной совокупности заданий. В классической теории ошибочный компонент измерений рассматривается как случайная погрешность неизвестного происхождения, что не позволяет в ее рамках исследовать надёжность результатов опроса в зависимости от таких источников погрешностей, как нестабильность результатов испытуемых, недостаточная внутренняя состоятельность тестовых результатов, несогласованность экспертов. Это, напротив, может быть оценено с помощью аппарата расширенной статистической GT–теории [1–6].

Другая современная широко используемая модель теории тестов – Item Response Theory (IRT) основывается на том, что вероятность правильного ответа испытуемого на задание теста рассматривается как функция, зависящая от латентных параметров: уровня подготовленности испытуемого и меры трудности задания [1–6]. Однако, результаты применения IRT оказываются чувствительными к нарушениям исходных предпосылок этой теории, а также к размеру выборки испытуемых. Для обеспечения достаточной точности требуется от 200 до 1000 испытуемых [1–6], что далеко не всегда бывает реализуемым в реальном учебном процессе. Классическая статистическая и расширенная статистическая теории оказываются более устойчивыми к нарушениям исходных положений.

С точки зрения методологии, все три упомянутые теории – классическую, расширенную статистическую и IRT трудно сравнивать, потому что они имеют различные предметы исследования, однако следует ожидать, что наиболее эффективным будет использование их в комплексе для решения отдельных задач оптимизации тестового контроля.

МЕТОДИКА АНАЛИЗА

В ходе данной работы была проведена оценка качества тестов и прогнозирование путей улучшения их качества, которая включала три взаимосвязанных блока: блок подготовки данных для анализа, блок моделирования обработки теста на основе классической и расширенной статистической теорий тестов и блок моделирования обработки теста согласно многопараметрической Item Response Theory.

Данными для анализа служили результаты тестовых испытаний студентов факультета компьютерных наук по курсам «Архитектура ЭВМ» и «Управление данными», проводимых с использованием средств системы электронного обучения Moodle [8]. Доступ к базам тестовых заданий и результатов тестирования, реализованным в MySQL, осуществлялся с использованием программных средств XAMP и Moodle. Моделирование оценки качества тестов проводилось средствами Microsoft Excel.

Параметры выборок для анализа тестовых сценариев приведены в (табл. 1).

При моделировании выгруженные из Moodle исходные баллы преобразовывались к приведенной матрице результатов [1, 5, 7]. Полученная матрица использовалась для расчета статистических параметров отдельных вопросов и теста в целом по расширенной статистической (Generalizability Theory), и современной параметрической теории Item Response Theory.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Оценка качества тестов с позиций Generalizability Theory

На первом этапе был проведен анализ факторов, влияющих на качество тестового задания, включавший статистическую оценку отдельных вопросов и тестов в целом средствами, заложенными в системе Moodle [2]. Были получены следующие характеристики для заданий и тестов: Индекс легкости, Стандартное отклонение баллов, параметр Случайности угадывания оценки, Предполагаемый вес вопросов, Эффективный вес вопросов, Коэффициент дифференциации и Эффективность дифференциации.

Мера легкости задания p_j ($j = 1, 2, \ldots, m$: m – число вопросов, n – число испытуемых, x_{ij} – оценка успешности выполнения j-го задания i-м опрашиваемым) позволяет выделить вопросы, не соответствующие уровню подготовки испытуемых:

$$p_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}.$$

Чем больше величина коэффициента p_j, тем большая часть испытуемых успешно справляется с заданием j (рис. 1).

Сложность заданий теста должна соответствовать уровню подготовки испытуемых. Тест в целом должен включать в себя комплекс заданий различной сложности – от легких до трудных [3, 6–8, 10]. Слишком простые задания, на которые правильно отвечают все испытуемые ($p_j = 1$), и слишком сложные задания, на которые не может ответить никто из опрашиваемых ($p_j = 0$), не обладают способностью дифференцировать испытуемых по уровню их подготовки и должны из теста исключаться.

Оценка индекса лёгкости позволила выявить в тесте неинформативные вопросы. В случае теста Контроль № 1 в курсе «Управление данными» имеют место легкие вопросы: вопрос 14 (0,90) и вопрос 17 (0,95), Полученные данные для теста Контроль № 1 курса «Архитектура ЭВМ», свидетельствует о том, что в используемой базе тестовых заданий присутствует 23,26 % (70 вопросов) очень лёгких ($p_j = 1,00$) и 17 вопросов (5,65 %) очень сложных – «никем не решаемых», с индексом легкости ниже 0,10.

Таблица 1

Параметры выборок для анализа

Курс	Время выполнения	Вопросов/ категорий в банке	Вопросов в тесте	Число первых попыток
«Архитектура ЭВМ» тематический тест №1	45 мин	301/34	34	Опрос 1 - 91 Опрос 2 - 230
«Архитектура ЭВМ» тематический тест №2	45 мин	89/29	29	Опрос 1 - 116 Опрос 2 - 208
«Архитектура ЭВМ» тематический тест №3	45 мин	120/43	43	Опрос 1 - 120 Опрос 2 - 102
«Архитектура ЭВМ» итоговый тест	45 мин	258/38	38	Опрос 1 - 79
«Управление данными», тематический тест №1	50 мин	132/50	52	Опрос 1 – 227
«Управление данными» тематический тест №2	50 мин	260/50	50	Опрос 1 – 223
«Управление данными», итоговый тест	50 мин	247/50	50	Опрос 1 – 16

Стандартное отклонение результатов испытуемых по j-му заданию теста ($j = 1, \ldots, m$) вычисляется по формуле:

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_i \right)^2}{n-1}},$$

где x_{ij} – оценка успешности выполнения j-го задания, выполненного i-м испытуемым (рис. 2).

Анализ величин стандартного отклонения оценки для каждого вопроса позволяет выявить ее вклад в дифференцирующую способность теста. Для большинства тестовых вопросов в исследованных тестовых заданиях оно имеет значение большее 0,30, что, в соответствии с требованиями педагогической теории измерений [1, 5, 6, 7, 9], является хорошим показателем их дифференцирующей способности (рис. 2). Задания, для которых это значение меньше 0,30, дифференцирующей способностью не обладают и должны быть переработаны. Например, в случае теста Контроль № 1 курса «Управление данными», заданий, требующих пересмотра по величине стандартного отклонения всего одно – № 14 (0,29). В случае теста Контроль № 1 курса «Архитектура ЭВМ» заданий, имеющих стан-

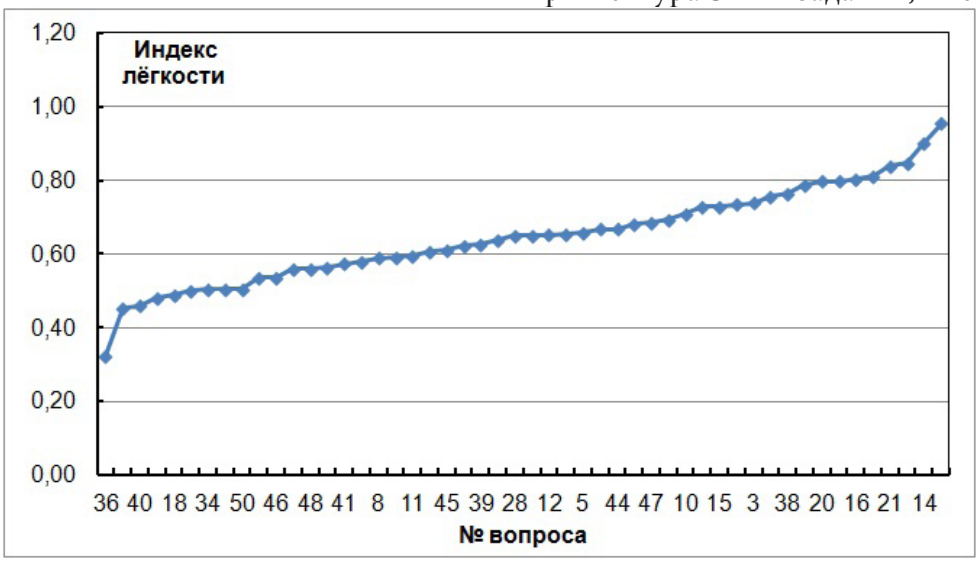

Рис. 1. Значения индекса лёгкости заданий теста в выборке для теста Контроль № 2 курса «Управление данными»

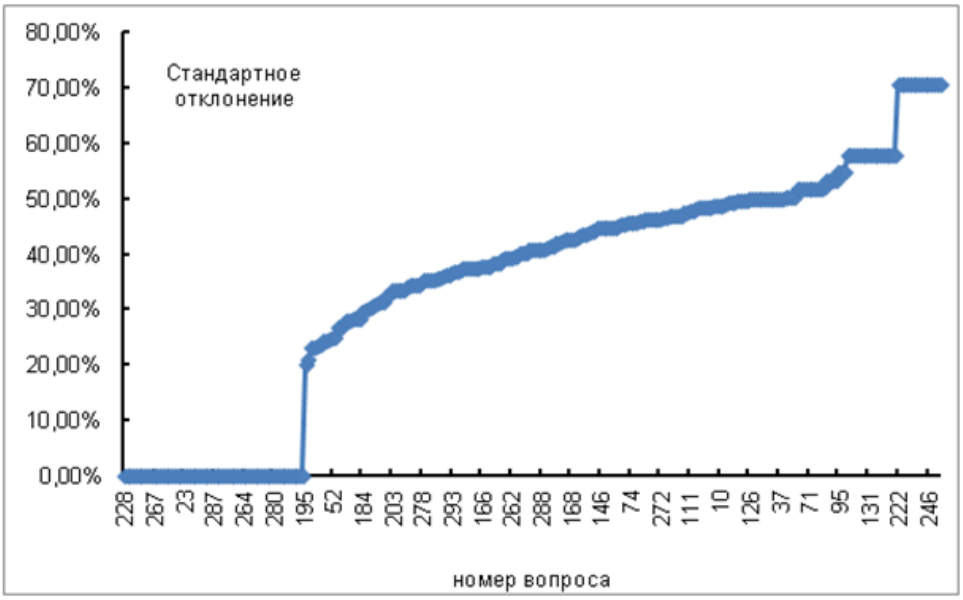

Рис. 2. Стандартное отклонение результатов выполнения заданий от номера вопроса в выборке теста Контроль № 1 курса курса «Архитектура ЭВМ»

дартное отклонение оценки ниже 0,27–82, и они относятся к одной трети категорий банка вопросов (рис. 2).

Важной статистической характеристикой дифференцирующей способности тестовых заданий, которую можно получить средствами Moodle, является **Коэффициент Дифференциации** – DE_j $(j=1,2,\ldots,m)$ – индикатор корреляции между оценкой за этот вопрос и оценкой за тест в целом.

$$DE_j = 100\frac{C(x_j,T)}{C_{\max}(x_j,T)},$$

$$C(x_j,T) = \frac{1}{S-1}\sum_{s\in S}\left(x_j(s)-\bar{x}_j\right)\left(T_S-\bar{T}_S\right).$$

$\bar{x}_j$ – среднее значение баллов, всех испытуемых за выполнение j-го задания, на которое получено S ответов; T_s – средняя оценка за задание; $\bar{T}_s$ – средняя оценка за тест. Этот показатель принимает значения между –1 и +1 и является мерой способности конкретного задания разделять сильных и слабых испытуемых. Положительные значения соответствуют заданиям, которые действительно разделяют «сильных» и «слабых» студентов, в то время как отрицательное значение DE_j свидетельствует о том, что плохо подготовленные студенты отвечают на данное задание в среднем лучше, чем хорошо подготовленные. Очевидно, что такие задания не являются тестовыми, так как не способны адекватно разделять испытуемых по уровню их подготовленности, и их следует отбраковывать. Считается, что задание обладает достаточной дифференцирующей способностью, если коэффициент дифференциации имеет значение больше или равное 0,30 [2, 6, 7].

Анализ величины DE_j для рассматриваемых тестов позволил выявить задания, не обладающие достаточной дифференцирующей способностью ($DE_j < 0,30$). В тесте Контроль № 1 курса «Архитектура ЭВМ» тестовых вопросов, не удовлетворяющих требованиям, оказалось 9,3 %, более того, в 15 случаях значение этого коэффициента отрицательное, что свидетельствует необходимости пересмотра этих вопросов. В тесте Контроль № 1 курса «Управление данными» всего 8 тестовых вопросов не удовлетворяют требованиям $DE_j < 0,30$ (рис.3).

На следующем этапе моделирования из собрания тестовых заданий удалялись задания слишком легкие ($p_j > 0,9$) и слишком трудные ($p_j < 0,2$), либо имеющие низкую дифференцирующую способность ($DE_j < 0,30$) или низкое стандартное отклонение оценки ($\sigma_j < 30$). Для измененных списков заданий вновь проводилась статистическая оценка для определения влияния этих факторов на качество теста.

Важнейшей характеристикой теста в целом является его надежность, характеризующая

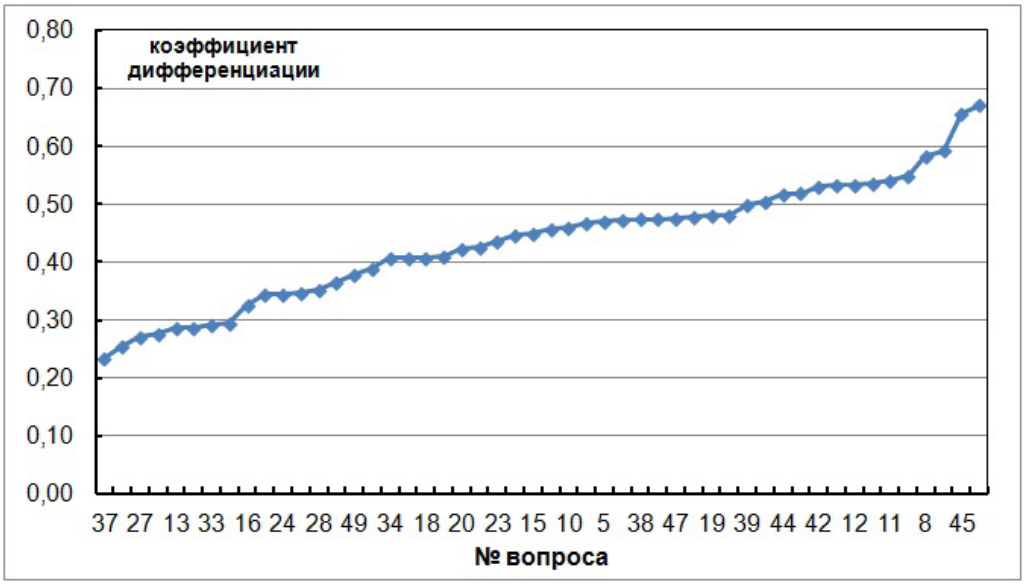

Рис. 3. Значение эффективного коэффициента дифференциации заданий теста Контроль № 1 курса «Управление данными»

воспроизводимость результатов тестирования, и их точность. Коэффициент надежности – это корреляционный коэффициент, показывающий степень совпадения результатов сеансов тестирования, проводимых в одинаковых условиях, одним и тем же тестом на разных выборках опрашиваемых. Надежность теста зависит от ошибки измерений [1, 5, 6, 9]. Когда ошибка (SE) отсутствует, коэффициент надежности равен единице. Если измеренный тестовый балл полностью обусловлен ошибкой измерения, то надежность теста равна нулю. Если известна ошибка SE, то при дальнейшем использовании теста можно считать, что студент выполнил тест успешно, если его оценка лежит в пределах ± SE, где SE – ошибка, найденная в предыдущих тестированиях. Согласно теории педагогических измерений считается, что ошибка при использовании качественного теста не должна превышать 8 % [1, 6].

На примере тестового задания Контроль № 1 курса «Архитектура ЭВМ» было проведено моделирование тестового сценария с целью установления факторов, определяющих его качество и прогнозирования его улучшения. Моделирование теста включало этап редактирования тестового сценария с последующей оценкой всех статистических параметров [2]. Редактирование теста заключалось в «исключении» из сценария тестирования «некачественных» вопросов, выявленных по описанным ранее параметрам (по эффективному весу вопроса, по величине стандартного отклонения, по эффективному коэффициенту дифференциации) путем присвоения максимально возможной оценке по этим заданиям значения «0». Оценка качества теста в целом проведена с использованием встроенного статистического модуля системы электронного обучения Moodle (табл. 2). В качестве критерия надёжности теста использовался коэффициент внутренней согласованности теста (коэффициент Кронбаха) [2].

Было установлено, что исключение из теста «забракованных» вопросов приводит к увеличению внутренней согласованности теста и уменьшению стандартной ошибки (табл. 2). Следует отметить, что в рассматриваемом случае, основной вклад в увеличение надёжности теста вносит исключение из теста вопросов, «выбракованных» по низкому коэффициенту дифференциации. Влияние двух других анализируемых факторов (индекса лёгкости и стандартного отклонения) проявилось не столь значительно. Это может быть обусловлено структурой теста, в частности, с большим разбросом частоты использования вопросов из банка.

Статистический анализ тестового задания на выборке в 215 вопросов показал увеличение критерия надёжности теста в Moodle до

Таблица 2

Статистические характеристики сценария теста Контроль № 1 курса «Архитектура ЭВМ» в системе Moodle

Название теста	Количество полных оцененых первых попыток	Общее количество полностью оцененных попыток	Средняя оценка для первой попытки	Средняя оценка по всем попыткам	Медиана оценки (для первой попытки)	Стандартное отклонение (для первой попытки)	Асимметрия (для первой попытки)	Эксцесс (для первой попытки)	Коэфф. внутренней согласованности (надежности) теста (для первой попытки)	Error ratio (для первой попытки)	Стандартная ошибка (для первой попытки)
Тест атт 1	91	97,00	73,77%	72,07%	77,03%	19,34%	-0,741	-0,099	89,53%	32,35%	6,26%
после удаления "некачественных" вопросов											
Тест атт 1	91	97,00	73.32%	71.65%	77.03%	19.97%	-0,7616	-0,1389	90.36%	31.04%	6.20%
увеличение числа тестируемых											
Тест атт 1	215	234	72,72%	70,91%	77,10%	19,79%	-0,7782	-0,2324	89,93%	31,73%	6,28%

0,8993 и стандартной ошибки определения оценки, получаемой студентом до 0,0628 при увеличении числа оценок в выборке (табл. 2). Полученный результат хорошо согласуется с имеющейся точкой зрения, что увеличение размера выборки приводит к увеличению надежности теста [1, 5, 6].

Определение надежности реальных тестов осуществляют сопоставлением индивидуальных баллов разных сеансов тестирования с использованием коэффициент корреляции Пирсона для результатов этих сеансов [3]. Из-за различия количества первых попыток в анализируемых независимых тестах, размеры выборок оценок для тестов, сравниваемых по критерию Пирсона, приводились к размеру меньшей выборки (91 первая попытка) путем случайного отбора значений из большей исходной матрицы результатов тестирования. Оценка двух независимых попыток теста Контроль № 1 курса «Архитектура ЭВМ» на двух независимых выборках студентов показала хорошее согласование их результатов. Критерий Пирсона для двух независимых выборок составил 0,9899 (табл. 3), что свидетельствует о высокой надежности тестового сценария.

Считается, что определение надежности теста необходимо выполнять на специально подобранной выборке испытуемых, репрезентативно представляющей всю генеральную совокупность, которая должна быть достаточно большой – 200–300 человек. Размер выборки определяет надежность теста [1, 3, 6]. Анализируя полученные результаты для нашего случая, также можно заключить, что увеличение числа первых попыток (числа испытуемых), увеличивает надёжность теста и уменьшает стандартную ошибку оценки, получаемой студентом. Более того, коэффициент корреляции оценок двух независимых сеансов тестирования показал применимость рассчитываемой в Moodle оценки качества тестов для выборки испытуемых 60–100 человек.

В случае невозможности использования повторного или независимого тестирования для оценки качества тестового сценария можно также использовать результаты одного сеанса тестирования путем расщепления его результатов на две независимые группы и нахождения корреляции между ними (split-half method) [3, 6]. Надежность теста методом расщепления определялась с использованием результатов однократного тестирования (91 попытка, Контроль № 1, «Архитектура ЭВМ»). Откорректированную тестовую матрицу для 34 вопросов разбивали на две части, состоящие из заданий с четными и нечетными номерами. Коэффициент корреляции Пирсона $r_{1/2}$ между двумя совокупностями суммарных баллов результатов этих выборок составил 0,867, что свидетельствует о хорошей надежности теста (табл. 3). Так как для определения надежности использовались половинки теста, то полученное значение критерия является заниженным (из-за уменьшения числа вопросов в тесте) и корректируется использованием соотношения Спирмена-Брауна [1, 6]:

$$\rho = \frac{2r_{1/2}}{1+r_{1/2}}.$$

Здесь $r_{1/2}$ – коэффициент надежности по половинкам расщепленного теста, а ρ – скорректированный коэффициент надежности. В анализируемом случае этот коэффициент составил 0,929, что говорит об удовлетворительной надежности теста (больше 0,7) [1, 5, 6].

Другой способ оценки надежности расщепленного теста основан на формуле Рюлона:

$$\rho = 1 - \frac{s_d^2}{s_y^2},$$

Таблица 3

Коэффициенты надежности теста Контроль № 1 курса «Архитектура ЭВМ»

Критерий	**Значение**
Коэфф. внутренней согласованности вMoodle (критерий Кохрена)	0,895
Расщепления (Пирсона)	0,867
Спирмена-Брауна	0,929
С исп. среднего коэфф. корреляции вопросов (Рюлона)	0,979
Критерий Пирсона по двум независимым тестированиям	0,990

где s_y^2 – дисперсия суммарных баллов результата, а s_d^2 – дисперсия разностей между результатами каждого испытуемого по обеим половинам теста.

Значение критерия Рюлона, составившее 0,979, хорошо согласуется со значениями коэффициента надежности, оцененными по другим методами (табл. 3). Таким образом, значения коэффициентов надежности, полученные разными способами, хорошо согласуются между собой (> 0.80), что свидетельствует о высокой надежности теста. Отметим, также, что автоматически рассчитываемый в Moodle для оценки надежности теста коэффициент внутренней согласованности теста может служить адекватной характеристикой качества используемого тестового сценария.

Проведенный анализ выборок первых попыток ответов тестов в курсе «Управление данными» (табл. 4) продемонстрировал высокую надежность тестовых сценариев для выборок размером ~ 200 и некорректность оценки в случае выборки размером в 16 испытуемых, объясняемой ее малым размером.

Оценка качества тестов с позиций Item Response Theory

Основным подходом в IRT является установление связи между двумя множествами значений латентных параметров. Первое множество составляют значения латентного параметра, определяющего уровень подготовленности испытуемых θ_i (i – номер испытуемого, изменяющийся в интервале от 1 до N, а N – количество испытуемых). Второе множество составляют значения латентного параметра, характеризующего трудность j-го задания β_j. Индекс j меняется в пределах от 1 до M, где M – количество заданий в тесте. Предполагается, что уровень подготовленности испытуемого θ_i и уровень трудности задания β_i оцениваются по одной шкале и измеряются в одних и тех же единицах – логитах [1, 3, 6]:

логит уровня подготовки i-го ученика $\theta_i = \overline{\beta} + Y\theta_i^0,\ \theta_i^0 = \ln\dfrac{p_i}{q_i}$,

логит трудности j-го задания $\beta_j = \overline{\theta} + X\beta_j^0,\ \beta_j^0 = \ln\dfrac{q_j}{p_j}$,

где p_i и q_i – доли, соответственно, правильных и неправильных ответов i-го ученика на задания теста; X и Y – поправочные коэффициенты [3], учитывающие дисперсии логитов по выборке.

Согласно параметрической модели Г. Раша, успешность решения заданий теста имеет вероятностный характер и измеряется числом P ($P \in [0,1]$). Принято считать, что вероятность того, что определенный участник тестирования, верно, решит определенное задание, представляет собой функцию успеха [1, 3–6], зависящую от двух аргументов (от разности $\theta_i - \beta_i$):

$$P_j(\theta) = \frac{\exp(1.7(\theta - \beta_j))}{1 + \exp(1.7(\theta - \beta_j))},$$

Таблица 4

Статистические характеристики тестовых заданий курса «Управление данными»

Название курса	Идентификатор	Количество полных оцененых первых попыток	Средняя оценка для первой попытки	Медиана оценки (для первой попытки)	Стандартное отклонение (для первой попытки)	Асимметрия (для первой попытки)	Эксцесс (для первой попытки)	Коэффициент внутренней согласованности	Error ratio (для первой попытки)	Стандартная ошибка (для первой попытки)	Банк вопросов	Число вопросов в тесте	Время тестирования
Управление данными	Тест_Атт1	227	76%	78%	16%	-0,72	0,19	90%	31%	5%	132	52	50
Управление данными	Тест Атт2 11	223	64,80%	66,60%	21,10%	-0,273	-0,684	92,70%	27,00%	5,70%	260	50	50
Управление данными	Тест экзамен	16	38,84%	41,35%	9,57%	-0,2452	-1,1141	47,89%	72,19%	6,91%	247	40	50

$$P_i(\beta) = \frac{\exp(1.7(\theta_i - \beta))}{1 + \exp(1.7(\theta_i - \beta))}.$$

Оценку качества тестового опроса проводят на основе анализа характеристических кривых отдельных заданий (огив), отражающих взаимосвязь между значениями независимой переменной θ (уровнем подготовленности) и успешностью выполнения этого задания $P_j(\theta)$. Точке перегиба характеристической кривой соответствуют значения $\theta = \beta_j$; $P_j = 0,5$, соответствующие тому, что испытуемый с уровнем подготовки θ равным трудности j-го задания теста, ответит на него правильно с вероятностью 0,5. Для испытуемых с уровнями знаний намного большими β_j, вероятность правильного ответа стремится к единице. Если же θ расположено достаточно далеко от значения $\theta = \beta_j$ и ниже 0,5, то вероятность правильного выполнения j-го задания теста близка к нулю.

На рис. 4 представлены результаты моделирования функции успеха по результатам оценки уровня подготовки испытуемых и трудности задания в критериально-ориентированном тесте, предназначенном для выявления степени усвоения студентами темы курса «Управление данными» (Контроль № 2) (рис. 4.а). Полученные зависимости позволили оценить трудность отдельных зданий и теста в целом. В анализируемом случае наиболее трудным по отношению к остальным оказалось задание, имеющее уровень трудности 1,98 (№ 36), а самым легким № 17 с уровнем трудности –2,53. В то же время, оказалось, что огивы неравномерно распределены по интервалу подготовленности студентов и иллюстрируют невысокую общую сложность тестового сценария (кривые смещены в область малых значений подготовленности) (рис. 4.а).

Так как в однопараметрической модели Г. Раша все задания обладают одинаковой дифференцирующей способностью, то для полного решения задачи отбора наиболее эффективных заданий при конструировании теста может оказаться недостаточным анализа в рамках этой модели. Для учета дифференцирующей способности заданий, А. Бирнбаумом в функцию успеха был введен дополнительный параметр a_j, учитывающий дифференцирующую способность j-ого задания теста, отражающийся на крутизне огивы [1, 3–6]:

$$P_j\left\{x_{ij} = 1 \middle| \beta_j\right\} = \left\{1 + \exp\left[-1,7 a_j(\theta - \beta_j)\right]\right\}^{-1},$$

$$a_j = \frac{(r_{bis})_j}{\sqrt{1 - \left[(r_{bis})_j\right]^2}},$$

где $(r_{bis})_j$ – бисериальный коэффициент корреляции j-го задания [1, 4].

Анализ характеристических кривых заданий одинаковой трудности, но разной крутизны позволяет отобрать оптимальные задания и определить границы интервала для значений параметра a_j (рис. 4.б). При этом проводится уменьшение длины теста за счет удаления части заданий равной трудности с более низкой дифференцирующей способно-

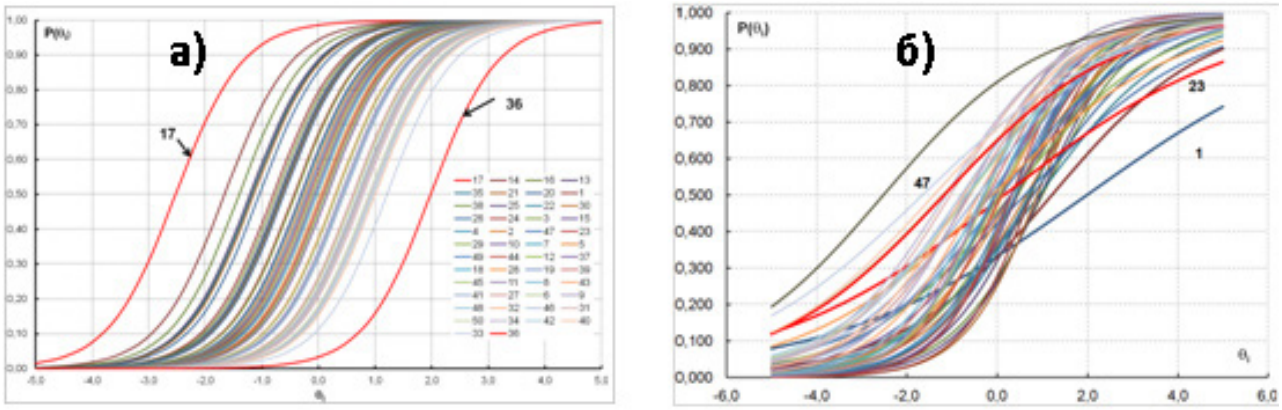

Рис. 4. Характеристические кривые тестовых заданий для теста Контроль № 2 курса «Управление данными»: а) однопараметрическая модель Г. Раша; б) двухпараметрическая модель А. Бирнбаума

стью. Наиболее эффективным считается задание с наибольшим значением параметра a_j. Теоретически значения параметра a_j могут изменяться в интервале $(-\infty,+\infty)$, но не все задания с параметром a_j из этого интервала можно включать в тест. На практике рекомендуется оставлять задания со значениями a_j, в интервале (0,5…2,5) [1,3–6]. Значение $a_j = 1$ соответствует однопараметрической модели Раша. Анализ полученных зависимостей позволил выявить в анализируемом тесте задания, имеющие относительно малую дифференцирующую способность: № 1, 23 и 47 ($a_j = 0,427$; 0,433; 0,574 соответственно), требующие корректировки.

Согласно А. Бирнбауму [2–6], количество информации, обеспеченное j-м заданием теста в данной точке θ_i – это величина, обратно пропорциональная стандартной ошибке измерения значения θ_i с помощью j-го задания. Для количественного описания информации, соответствующей заданию используется информационная функция $I_j(\theta)$:

$$I_j(\Theta) = 2.86\frac{\exp(1.7(\Theta-\beta_j))}{(1+\exp(1.7(\Theta-\beta_j)))^2}.$$

Информационные функции вопросов теста обладают свойством аддитивности, что позволяет построить информационную функцию всего теста (рис. 5). Информационная функция теста должна иметь один четко выраженный максимум. Если это не так, то тест нуждается в доработке путем добавления в него заданий с трудностями, соответствующими областям «провала» информационной функции теста.

В нашем случае для теста Контроль № 1 курса «Архитектура ЭВМ» информационная функция имеет четко выраженный максимум ($I_{макс}(0.110) = 26.78$), симметрична относительно значения $\theta_j = 0.110$, что свидетельствует о равномерном вкладе в результат теста вопросов различной сложности (рис. 6.а). Однако, в случае тестового задания Контроль № 2 курса «Управление данными» экспериментальные значения $I_j(\theta)$ сконцентрированы в области значений θ, отвечающих высокому уровню подготовленности студентов, что свидетельствует о преобладании в тесте легких заданий.

ЗАКЛЮЧЕНИЕ

С целью разработки сценария оценки качества тестов и тестовых заданий в процессе их создания и использования, был проведён сравнительный анализ методик оценки параметров надежности тестов и качества входящих в них вопросов с использованием моделей современной теории педагогических измерении: расширенной статистической (Generalizability Theory) и параметрической теории Item Response Theory. Анализ проведен по результатам тестирований студентов факультета компьютерных наук по дисциплинам «Управлении данными» и «Архитектура ЭВМ».

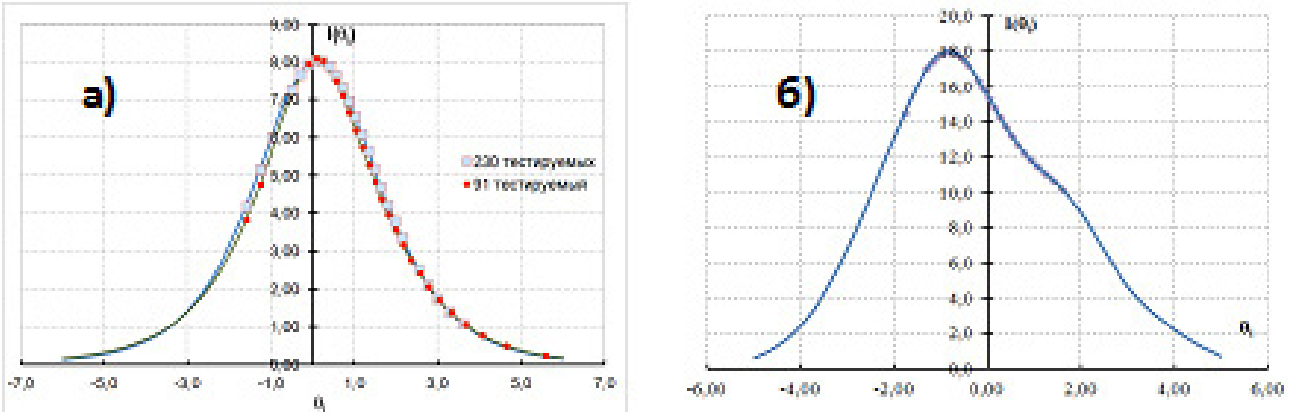

Рис. 5. Информационная функция по однопараметрической модели Г.Раша (линия) и экспериментальные значения (точки): а) теста Контроль № 1 курса «Архитектура ЭВМ» б) теста Контроль 2 курса «Управление данными»

Сопоставление различных критериев надежности тестового сценария показало высокое значение этой характеристики для всех исследованных тестов (> 0,80) и хорошее согласование разных подходов к оценке достоверности оценки за тест. Анализ подтвердил, что увеличение длины теста сопровождается повышением его надежности. Рассмотренные тестовые сценарии содержали от 29 до 52 вопросов. Это количество вопросов оказалось достаточным для надежной оценки знаний испытуемых.

Существенное влияние на точность оценки надежности теста оказывает размер выборки ответов, на основании которой определяется оценка за тест. Число первых попыток в анализируемой выборке менялось о 61 до 230, что обеспечило достаточную точность оценки результатов тестирования.

Оценка качества входящих в тесты вопросов, проведенная на основе результатов тестирований в системе электронного обучения Moodle с позиции расширенной статистической теории позволила выявить факторы, влияющие на надежность тестовых заданий. Уровень сложности вопроса, его дифференцирующая способность и величина стандартного отклонения оценки значительно влияют на надежность теста. Увеличение коэффициента дифференциации теста за счет введения разного веса вопросов приводит к уменьшению его согласованности, но снижает ошибку определения оценки. Использование модели Generalizability Theory может быть рекомендовано для оценки качества тестов с размером выборки первых попыток ответов более 40 и числом вопросов более 20.

Сопоставление оценок качества тестов, основанных на различных моделях современной теории педагогических измерений показало, что при разработке контрольно-измерительных материалов в тестовой форме необходима оценка качества создаваемого теста и его корректировка в процессе использования. При оценке качества теста следует проводить анализ, как качества вопросов, так и теста в целом. Необходимым, но недостаточным условием анализа качества тестовых материалов является использование математического аппарата классической теории. Корректная оценка качества вопросов возможна с привлечением современной параметрической модели Item Response Theory.

Расширенная статистическая теория (Generalizability Theory), реализованная в модуле статистического анализа среды Moodle, позволяет учитывать влияние длины теста, числа тестируемых студентов, а также свойств вопросов, таких как, сложность, стандартное отклонение оценки по вопросу, вероятность угадывания на надежность теста и на ошибку определения оценки. Представляется достаточным на первом этапе при создании тестового сценария в электронной образовательной среде Moodle проводить оценку его качества с помощью встроенного модуля статистического анализа. Это позволит создавать контрольно-измерительные материалы в тестовой форме, корректно оценивающие качество подготовки студентов, без необходимости проведения сложных статистических вычислений.

СПИСОК ЛИТЕРАТУРЫ

1. *Челышкова М. Б.* Теория и практика конструирования педагогических тестов. – Москва : «Логос», 2002. – 431 с.

2. Сайт MoodleDocs. – (http://docs.moodle.org/ru/)

3. *Ronald K. H.* Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development / K. Hambleton Ronald, l W. Jones Russel // Educational Measurement: Issues and Practice. – 1993. – P. 38–47.

4. *Гласс Дж., Стэнли Дж.* Статистические методы в педагогике и психологии / Дж. Гласс, Дж. Стэнли – М. : Прогресс, 1976. – 496 с.

5. *Майоров А. Н.* Теория и практика создания тестов для системы образования / А. Н. Майоров. – М. : Народное образование, 2000. – 352 с.

6. *Ким В. С.* Тестирование учебных достижений. – Уссурийск : Издательство УГПИ, 2007. – 214 с.

7. *Толстобров А. П., Коржик И. А.* Возможности анализа и повышения качества тестовых

заданий при использовании сетевой системы управления обучением MOODLE. – Вестник Воронежского государственного университета. Сер. Системный анализ и информационные технологии. – Воронеж, 2008. – № 2. – С. 100–106.

8. Официальный сайт центра электронных образовательных технологий ВГУ. – (http://www.moodle.vsu.ru)

9. Страница официального сайта программы XAMPP для выгрузки дистрибутива для ОС Windows. – (http://www.apachefriends.org/en/xampp-windows.html)

10. *Аванесов В. С.* Содержание тестов и тестовых заданий [Электронный ресурс] / В. С. Аванесов // Педагогические Измерения, 2007. – № 3. – (http://testolog.narod.ru/Theory63.html)

Протасова Ирина Валентиновна – к. х. н., доцент кафедры физической химии химического факультета Воронежского государственного университета. Тел.: (473) 220-85-38
E-mail: protasova@chem.vsu.ru

Толстобров Александр Павлович – к. т. н, доцент кафедры информационных систем факультета компьютерных наук, директор Центра электронных образовательных технологий. Тел.: +7(473) 255-56-46
E-mail: tap@vsu.ru

Коржик Илья Андреевич – директор Областного центра технического творчества учащихся. Тел.: +7(473) 220-74-68
E-mail: ikorzhik@gmail.com

Protasova Irina Valentinovna – Cand. Sci. (Chem.), Associate Professor of the Physical Chemistry Department, Voronezh State University. Tel.: (473) 220-85-38
E-mail: protasova@chem.vsu.ru

Tolstobrov Alexander Pavlovich – Candidate of Technical Sciences, Associate Professor of the dept. Information Systems, Computer Sciences Faculty, Director of the Center of Electronic Educational Technologies, Voronezh State University. Tel.: +7(473) 255-56-46
E-mail: tap@vsu.ru

Korzhik Iliy Andreevich – Director of the Regional Centre of Technical Creativity.
Tel.: +7(473) 220-74-68
E-mail: ikorzhik@gmail.com