Бузун, Д.Н. Компьютерные дидактические тесты: оценка качества / Д.Н. Бузун // Информационное обеспечение исторического образования: Сб. ст. / Под. ред. В. Н. Сидорцова, А. Н. Нечухрина, Е. Н. Балыкиной. – Минск: БГУ; Гродно: ГрГУ, 2003. – С. 76–86. (Педагогические аспекты исторической информатики; Вып. 3).

**Д. Н. Бузун**
(Минск, БГУ)

# КОМПЬЮТЕРНЫЕ ДИДАКТИЧЕСКИЕ ТЕСТЫ: ОЦЕНКА КАЧЕСТВА

В настоящее время, в период перехода системы высшего образования на конвейерный метод производства, введения системы дистанционного образования особенно актуально встает вопрос о применении более эффективных методов контроля и оценки, чем традиционная письменная работа или устный экзамен [1]. Поэтому тестирование, как одна из эффективных форм оценки знаний учащихся активно внедряется в учебный процесс [2].

Первые стандартизированные тесты появились в США еще в начале прошлого века. В 1926 г. Совет колледжей принял тест SAT. Им же разрабатывались тесты для квалифицированной и профессиональной оценки деятельности педагога. С 1947 г. существует Служба тестирования (Educational Testing Service), которая считается одним из представительных научно-исследовательских центров. К 1961 г. только в США было создано 2126 стандартизированных тестов [3].

В РФ существует несколько центров, в которых профессионально разрабатывают тестовые методики, а также сами тесты по социально-гуманитарным дисциплинам. Среди них — Центр оценки качества образования Института общего среднего образования РАО, Центр тестирования выпускников общеобразовательных учреждений РФ, Центр психологического и профессионального тестирования МГУ, Лаборатория аттестационных технологий Московского института повышения квалификации работников образования (МИПКРО), НИИ образовательных процессов и программ СГУ и др. [3]

Центром тестирования МГУ «Гуманитарные технологии» в течение последнего времени создана и развивается Интернет-ориентированная система компьютерного тестирования «Телетестинг». Тестирование проводится по истории, русскому языку и др. В 1997–1999 гг. с использованием этой системы были проведены 3 всероссийские компьютерные олимпиады с одноименным названием, в которых приняли участие почти 16 000 человек из 70 различных городов РФ и ближнего зарубежья [4].

С 1999 г. действует Всероссийский центр тестирования. В том же году 40 000 абитуриентов стали студентами на основа-



нии сертификатов, полученных в Центре после прохождения тестирования [2].

По данным исследований, в Беларуси на данный момент не создано ни одного стандартизированного теста, однако, широко используются различного уровня педагогические тесты в школах и вузах страны [5, 6].

В статье ставится целью сопоставить различные варианты расчета основных показателей дидактических тестов, применяемые различными школами тестологии. Анализ тестовых заданий посредством математических методов позволяет получить информацию об их скрытых дефектах, которые не удается выявить с помощью экспертных методов.

К основным показателям качества тестовых заданий относятся:

- сложность,
- надежность,
- различительная способность,
- валидность.

**Сложность $p_j$** тестового задания **j** равна доле испытуемых, правильно ответивших на это задание:

$$p_j = \frac{n_j}{n}.$$

где

$n_j$ — число правильных ответов на j-задание;

**n** — общее число испытуемых, отвечавших на **j-е** задание.

Сложность всего теста можно рассчитать аналогично как долю правильных ответов на задание теста. Допустимый диапазон значения $p_j$ — 0,4–0,7. Очевидно, что задания с нулевой или стопроцентной сложностью должны быть исключены из тестового набора (такие задания не дифференцируют учащихся по уровню подготовки) [7].

Поскольку сложность тестового задания является случайной величиной, то имеет смысл говорить не только об оценке ее значения $p_j$, но и о доверительном интервале, соответствующем заданной доверительной вероятности $P_{дов}$. Границы доверительного интервала сложности ($P_{j\ min}$, $P_{j\ max}$) тестового задания можно рассчитать по формуле:



$$P_{j_{min}} = P_j - \frac{S_j t_{1-\alpha/2}}{\sqrt{n_j}}.$$

$$P_{j_{max}} = p_j + \frac{S_j t_{1-\alpha/2}}{\sqrt{n_j}}.$$

где

$p_j$ — оценка доверительной вероятности;

$n_j$ — общее число испытуемых, отвечавших на **j-е** задание;

$S_j$ — оценка стандартного отклонения **j-го** задания;

$t_{1-\alpha/2}$ — квантиль распределения Стьюдента;

$\alpha=1-P_{дов}$ — уровень значимости.

**Надежность** (*reliability*) — характеристика качества тестов, отражающая точность педагогических измерений, степень постоянства, стабильности, устойчивости результатов тестирования. Надежным считается тест, который дает постоянные результаты, оценки при повторных предъявлениях.

Коэффициент надежности теста **K** определяется как отношение:

$$K = S_t^2 / S_x^2.$$

где

$S_t^2$ — дисперсия истинной компоненты;

$S_x^2$ — дисперсия измеренных тестовых баллов.

Коэффициент надежности должен интерпретироваться не только как характеристика самого теста, но и как выборки аттестуемых, т. е. он зависит от уровня знаний, полученных в конкретном учебном заведении.

Для расчетов надежности по внутренней согласованности используются, как правило, статистические формулы. Одним из совершенных показателей для расчета надежности тестов, по мнению большинства специалистов, является коэффициент $\alpha$ «*альфа*»:

$$\alpha = \frac{k}{k-1}\left[1-\frac{\sum \delta_i^2}{\delta_y^2}\right].$$

где

**k** — количество задний;

$\sum \delta_i^2$ — сумма стандартных отклонений для заданий;

$\delta_y^2$ — квадрат стандартного отклонения для всего теста.

В эту формулу входят квадрат стандартного отклонения для всего теста, и чем он выше, тем больше коэффициент надежности, то есть чем больше дисперсия всего теста, тем он надежнее, и чем меньше сумма квадратов стандартных отклонений для каждого из заданий, тем больше значение коэффициента [3. С. 207].

В расчетах надежности часто применяется формула Кьюдера-Ричардсон, которая является частным случаем альфа Кронбаха для дихотомической оценки:

$$Kr_{20} = \frac{m}{m-1} \times (1 - \frac{\sum p_j \times q_j}{S_x^2}).$$

где

**m** — число заданий теста;

$p_j$ — сложность **j-го** задания;

$q_j = 1 - p_j$;

$S_x$ — стандартное отклонение суммарных индивидуальных баллов.



Допустимый диапазон изменения коэффициента надежности колеблется от 0,7 и выше [7,8].

П. Ж. Рюлон разработал формулу определения надежности методом расщепления [9. С.197], [10. С.22]:

$$r_{1l} = 1 - \frac{\sigma_d^2}{\sigma_x^2}.$$

где

$\sigma_d^2$ — дисперсия разностей между результатами каждого испытуемого по обеим половинам теста;

$\sigma_x^2$ — дисперсия суммарных баллов.

Для оценки статистической значимости коэффициента $\alpha$ для проверки надежности теста с 50-х гг. используют в тестологии формулу Дж. Китса [9. С.198]:

$$\chi_{n-1}^2 = \frac{k(n-1)}{k(1-\alpha)+\alpha}.$$

где

$\chi_{n-1}^2$ — эмпирическое значение статистики (кси-квадрат);

**n-1** — степени свободы;

**k** — количество пунктов;

**n** — количество испытуемых;

$\alpha$ — надежность.

Для получения сведений о надежности целого теста используется формула Спирмена-Брауна [9. С.197]:

$$r_{xx} = \frac{2r_x}{1-r_x}.$$



где

$r_x$ — эмпирически рассчитанная корреляция для половин;

$r_{xx}$ — надежность целого теста.

В дидактическом тестировании этот метод определения надежности подходит для сравнительно гомогенных по форме тестов, содержащих задания примерно одного уровня сложности [9],[8. С. 113]. Надежность теста обычно рассматривают на уровне всего тестового набора. Поскольку надежность теста обычно определяется количеством индивидуальных заданий теста и их согласованностью, говорить о надежности отдельного задания сложно. Однако допустимо рассматривать вклад отдельного тестового задания в надежность всего теста и согласованность результатов ответов на данное задание с результатами всего теста. В соответствии с этим можно ввести индекс надежности отдельного тестового задания $K_{rj}$ (*item reliability index*):

$$K_{rj} = K_{bj} \times \sqrt{p_j \; q_j}.$$

где

$K_{bj}$ — значимость.

$P_j$ — сложность конкретного задания.

$q_j=1-p_j$.

Характерные интервалы в норме для этого значения от 0,4 и более.

Существует еще один метод расчета коэффициента надежности — по Гутману. В соответствии с ним вместо коэффициента корреляции вычисляется среднее квадратичное отклонение результатов по четным и нечетным вопросам, а также по тесту в целом:

$$r_g = 1 - \sum l/(N \times K).$$

$r_g$ — коэффициент надежности по Гутману;

$\sum l$ — сумма ошибок;

$N$ — число испытуемых;

$K$ — число заданий;



Желательно значение $r_g \geq 0{,}8$.

Уровень доверия к полученному результату зависит от точности проведения измерения и количества испытуемых [9. С. 204–205].

**Различительная способность** задания при разработке педагогических тестов является особенно важной характеристикой, так как от нее в значительной степени зависит валидность теста. Эта характеристика показывает, насколько эффективно тестовое задание различает учащихся, овладевших и не овладевших учебным материалом.

Существует немало показателей для различения способностей, некоторые из них весьма сложны для вычисления. Мы рассмотрим только несколько из них, наиболее простые и в то же время достаточно эффективные.

Эти показатели требуют для расчета две серии измерений — повторного тестирования одной группы учащихся или проведения теста на двух разных группах.

При разработке теста для одной или небольшого количества групп учащихся удобнее всего получить две серии измерений путем формирования контрастных групп. Преподаватель выбирает из группы студентов только тех учащихся, про которых он может определенно утверждать, что они овладели или не овладели учебным материалом. Овладевшие материалом составляют «высокую» контрастную группу, а не овладевшие — «низкую». Учащиеся, находящиеся в промежуточном положении, не включаются в контрастные группы. Важно, чтобы эти группы были, по возможности, эквивалентны по составу. Это значит, что в них в одинаковой пропорции должны быть представлены испытуемые разных возрастов, обоего пола, личного уровня одаренности.

К сожалению, метод контрастных групп не может использоваться, если подавляющее большинство учащихся твердо овладели или совсем не овладели учебным материалом. Тогда для получения двух серий измерений приходится прибегать к методам, более сложным по организации. Можно протестировать одну и ту же группу учащихся до и после обучения или протестировать две группы (эквивалентные по составу, подобно контрастным), одна из которых прошла курс обучения, а вторая — нет.

Самый простой и широко известный показатель различительной способности задания по отношению к обучению **D** вычисляется как разность между долей испытуемых из «высокой» группы, правильно выполнившей задание, и долей испытуемых из «низкой» группы, также правильно выполнившей задание.



$$D=n_1/N_1 - n_2/N_2.$$

где $N_1$ и $N_2$ — количество испытуемых, попавших соответственно в «высокую» и «низкую» контрастные группы;

$n_1$ и $n_2$ — количество испытуемых, правильно выполнивших задание (соответственно из «высокой» и «низкой» групп).

Показатель может принимать значения от -1 до +1. **D = +1** означает, что задание обладает максимальной различающей способностью. **D = 0** означает, что задание совершенно не различает испытуемых, овладевших и не овладевших учебным материалом. Если **D =-1**, что встречается очень редко, то задание различает испытуемых, но правильно отвечает не овладевшие материалом, а овладевшие материалом отвечают неправильно. Существование таких заданий может свидетельствовать о своеобразной неадекватной структуре знаний у учащихся и должно быть причиной серьезных раздумий для педагога.

В последние годы был предложен другой показатель различительной способности задания **P(X)**, который считается более совершенным, чем **D** по ряду теоретических соображений. Показатель **P(X)** можно рассматривать как вероятность согласованности между результатом выполнения испытуемым задания и отнесением испытуемого к «высокой» или «низкой» контрастной группе. Он рассчитывается по формуле:

$$P(X)=n_1/N_1 + n_2/N_2.$$

где $N_1$ и $N_2$ — количество испытуемых, попавших соответственно в «высокую» и «низкую» контрастные группы;

$n_1$ — количество испытуемых из «высокой» группы, правильно выполнивших задание;

$n_2$ — количество испытуемых из «низкой» группы, неправильно выполнивших задание.

Наилучшие задания будут иметь значения **P(X)** равные единице. Минимальное значение показателя достигается в том случае, если между отнесением испытуемого к одной из групп и выполнением им задания не существует никакой связи [11].

**Валидность** (*validi*) — комплексная характеристика теста, отражающая обоснованность, значимость его результатов, адекватность теста целям измерения.

Наиболее распространенным способом нахождения теоретической валидности является конвергентная валидность, т. е.



сопоставление данной методики с авторитетными методиками и доказательством

значимых связей с ними. Сопоставление с методиками, имеющими другое теоретическое основание, и констатация отсутствия значимых связей с ними называется дискриминантной валидностью.

Другой вид валидности — прагматическая валидность — проверка методики с точки зрения и ее практической значимости, эффективности, полезности. Для проведения такой проверки, как правило, используются так называемые независимые внешние критерии. Среди них могут быть успеваемость, профессиональные достижения, достижения в разных видах деятельности, субъективные оценки (или самооценки).

Значимость тестового задания **j** отражает связь ответов на данное задание группы учащихся с индивидуальными баллами этой группы учащихся на **j-е** задание теста и индивидуальными баллами учащихся. Если принять во внимание тот факт, что результат ответа на **j-е** задание является дихотомической переменной, то можно получить следующее выражение для $\mathbf{K_{bj}}$:

$$K_{bj} = \frac{(B_{cpj} - B_{cp})}{S_x^2} \times \sqrt{\frac{p_j}{q_j}}.$$

где:

$\mathbf{B_{cpj}}$ — среднее значение индивидуальных баллов тех испытуемых, которые ответили на **j-е** задание правильно;

$\mathbf{B_{cp}}$ — среднее значение индивидуальных баллов всей выборки испытуемых;

$\mathbf{p_j}$ — сложность **j-го** задания.

$\mathbf{q_j = 1 - p_j}$

$\mathbf{S_x}$ — стандартное отклонение суммарных индивидуальных баллов.

Как обычно коэффициент корреляции, значимость **Kb** изменяется в пределах от — 1,00 до + 1,00. Приемлемыми считаются задания, у которых значимость больше или равна 0,3 [7].



Теоретические изыскания автора нашли свое практическое отражение в работе над расчетом основных показателей качества дидактического теста по одному из курсов, читаемых студентам исторического факультета Белгосуниверситета [12].

---

[1] Амзараков М. Б. Автоматическая генерация вариантов педагогического теста (http://www.ito.edu.ru/1999/II/6/6140.html).

[2] Кадневский В. М. История России в тестах: до начала XX века. М.: Школа – пресс, 1997.

[3] Майоров А. Н. Теория и практика создания тестов для системы образования. (Как выбирать, создавать и использовать тесты для целей образования). М.: "Интеллект-центр", 2001.

[4] Шмелев А. Г., Бельцер А. И., Ларионов А. Г. и др. Адаптивное тестирование знаний в системе «ТЕЛЕТЕСТИНГ» (http://www.teletesting.ru/tezadap.htm).

[5] См., напр., В. Н. Сидорцов, С. Б. Каун, С. Г. Мигуцкий Эксперимент по дистанционному обучению в системе WebCT // Информационный бюллетень Ассоциации "История и компьютер". №30. Материалы VIII конференции АИК. С. 238-239; Балыкина Е.Н. Шедевры иконописи Беларуси XII – XVIII вв. // Гісторыя: проблемы выкладання. Мн.: Адукацыя і выхаванне, 2000. № 1. С. 110-122, № 2. С.100-114; А. А. Яноўскі і інш. Экзаменацыйныя тэсты па гісторыі Беларусі. Мн.; Тэтра-сітэмс. 2002; Теория и история источниковедения: Учебно-метод. комплекс / С. Б. Каун, О. Л. Липницкая, С. Н. Ходин. Мн.: БГУ, 2001; Сергеенкова В. В. История России (1856-1917 гг.). Учебно-методический комплекс. Мн.: Издательский центр БГУ, 2002.

[6] См., напр., Белазаровіч В., Барыс С. Тэсты і заданні па гісторыі Беларусі // Беларускі гістарычны часопіс. 2000. №1. С. 84-89; Босы І. Тэставае апытанне як сродак аператыўнага кантролю на ўроках беларускай мовы і літаратуры // Роднае слова. 1998. №9. С. 125-131; Марчанка Г. Выкарыстанне тэстаў на ўроку (V клас) // Беларускі гістарычны часопіс. 1997. №1. С.128-132; Паноу С. Развіваючыя тэсты па гісторыі Беларусі (IX клас) // Гісторыя: праблема выкладання. 1998. №2. С. 53-70; Радзькоў А. М., Кравец А. У. Тэставыя методыкі ў навучальным працэсе: сучасны стан і перспектывы развіцця // Народная асвета. 2000. №8. С. 3-11; Радьков А. М., Кравец Е. В. Тестовые технологии в системе непрерывного образования: Методическое пособие. Могилев: МГУ им. А.А. Кулешова, 2001.

[7] Карпенко Д. С., Карпенко О. М., Шлихунова Е. Н Автоматизированная система мониторинга эффективности усвоения знаний и качества тестовых заданий // Инновации в образовании. 2001, №2. С.69-85.

[8] Аванесов В. С. Основы научной организации педагогического контроля в высшей школе. Пособие для слушателей Учебного центра Гособразования СССР. М.: МИСиС, 1989.

9 Михайлычев Е. А. Дидактическая тестология. М.: Народное образование, 2001.

[10] Аванесов В. С. Автореферат диссертации на соискание ученой степени доктора педагогических наук. СПб., 1994.

[11] Люсин Д. В. Основы разработки и применения критериально-ориентировочных педагогических тестов. Учебное пособие для слушателей повышения педагогической квалификации. М., 1993.

[12] Бузун Д. Н. Методы проведения анализа статистических результатов / Тезисы 3-й Всероссийской очно-заочной научно-практической конференции "Информационные технологии в управлении и учебном процессе вуза". Владивосток: ВГУЭС, 2002.