

МЕТОДЫ ОЦЕНКИ ЗНАНИЙ В СОВРЕМЕННЫХ ТЕСТИРУЮЩИХ СИСТЕМАХ

А.В. Шешуков

Оценка знаний в традиционном понимании рассматривается как определение итогового уровня подготовки обучаемого в рамках некоей предметной области после прохождения курса обучения. Существуют задачи контроля знаний и вне учебного процесса, которые используются в производственной среде. Основными из таких задач являются: аттестация персонала, проведение сертификационных экзаменов, отбор кандидатов на конкретные вакансии и т. п. Результатом оценки знаний, как правило, является определение некоторой величины в рамках шкалы оценивания.

В связи с чрезвычайной важностью точности оценивания знаний необходимо подходить к этому явлению как к процессу объективного измерения и результаты такого измерения обрабатывать стандартными математическими методами и сопровождать стандартными характеристиками точности. Весь процесс оценки знаний должен выполняться в рамках научно-обоснованной методологии, с практической апробацией, подтверждающей работоспособность теории.

Задачи исследования проблемы оценки знаний ставят несколько целей:

- максимизация объективности оценивания;
- увеличение эффективности процесса оценки;
- внедрение новых форм и способов оценки в современных методах обучения, аттестации, сертификации специалистов, и других сферах деятельности человека.

Научным направлением, лежащим в основе моделирования и параметризации процесса тестирования, являются две теории: так называемая «классическая теория тестирования» и «современная теория тестирования» получившая название в иностранной литературе название Item Response

Theory (IRT). В основе теории IRT лежит множество математических методов, основные из которых: теория вероятности, математическая статистика, дисперсионный и регрессионный анализ.

В традиционном или классическом тестировании уровень знаний, в конечном счете, определяется как доля правильных ответов [3]. Так в самом системном тестировании предполагается, что тестовые задания имеют одну и ту же трудность, и при таком же представлении определяется оптимальная длина теста и оценка надежности его результатов. В другой модификации традиционного тестирования разработчик сам приписывает тестовому заданию его трудность, например в 1, 2 или 3 балла. Набранный балл испытуемого определяется как число правильных ответов с учетом сложности заданий.

Бальная система оценивания (2, 3, 4 и 5), как уже было отмечено выше, также обладает существенным недостатком – она относительна и субъективна. Каждый преподаватель работает в своей ранговой шкале.

Фактически же во всех этих экспертных оценках трудности заданий определяются субъективно, что в конечном случае, отрицательно сказывается на общем уровне знаний.

Первый недостаток традиционного подхода – это большая зависимость тестового балла от ответов на число заданий, но и от состава, уровня подготовленности группы, что легче видеть на примере такой ситуации, как оценка трудности заданий. В заранее подготовленной группе, как правило, ответивших на задание существенно больше, чем в слабой группе. При этом всегда остается открытым главный вопрос – чему равняется истинное значение трудности заданий. Методы построения доверительных интервалов здесь неэффективны, потому что тест проверяется обычно не на случайном выборе, а на тех или иных студенческих группах, представляющих собой далеко неслучайно набранное множество испытуемых.

Второй недостаток классической теории тестов – это большая зависимость показателя надежности теста от вариативности тестовых баллов испытуемых в той или иной группе [2].

Традиционные тесты точнее оценивают средних студентов, чем самых лучших и самых худших. Это следствие общей ориентации создателей тестов на средний уровень. В тестах такого рода оценки каждого испытуемого рассчитываются относительно среднего балла, который, как было отмечено раньше, существенно меняется от группы к группе. Соответственно меняется и индивидуальный тестовый балл. Если испытуемый попадает в слабую группу, он оказывается в числе самых лучших; если (при том же результате) он попадает в сильную группу, то оказывается в числе худших. Вопросы такого рода возникают при профессиональном отборе, конкурсном приеме в вузы в случаях, если группы абитуриентов заранее не выравниваются; а это, как правило, так и бывает. Ошибки усугубляются, если принимается волевое решение отбирать половину (или другой процент)

каждой группы, независимо от показанных в них результатов. В этом случае судьба абитуриентов во многом зависит от группы и от установленной квоты приема из каждой группы.

Таким образом, можно отметить неустойчивость наиболее важных статистик, их взаимное влияние, существенно снижающее качество тестовых измерений. Эти статистики позволяют определить относительное положение каждого испытуемого в нормативной шкале. Специальные теоретические и экспериментальные исследования показали, что статистические поправки и ряд специальных приемов не позволяют в полной мере устранить отмеченные недостатки классической теории тестов.

Под современной теорией тестирования понимается распространенная на западе теория латентно структурного анализа – Item Response Theory (IRT), нацеленная на оценивание латентных качеств личности и параметров заданий теста на основе математических моделей измерений. В частности, в теории латентно-структурного анализа оцениваемые значения параметров рассматриваются как некоторые дискретные точки на оси латентной переменной, в то время как Item Response Theory распределения переменных предполагаются непрерывными [4].

Для измерения латентных переменных используется модель разработанная датским математиком Дж. Рашем [1]. Суть измерения латентной переменной на основе модели Раша состоит в следующем. Ответ i -го испытуемого по j -му тестовому заданию рассматривается как вероятностная функция взаимодействия двух основных факторов – уровня развития интересующего свойства (латентной переменной) θ_i , и уровня трудности тестового задания β_j . Вероятность правильного ответа испытуемого на вопрос теста тем ближе к единице, чем выше θ_i , и чем ниже β_j . Это условие компактно записывается как функция от разности двух параметров

$$P\{X_{ij} = 1 | \theta_i, \beta_j\} = \varphi(\theta_i - \beta_j). \quad (1)$$

Эту функцию принято называть функцией успеха. Из выражения (1) ясно, что высокая вероятность правильного ответа обеспечивается либо высокими значениями θ_i либо низкими β_j . При обратном соотношении, вероятность успеха низка и приближается к нулю.

Если вид функции успеха известен, то по результатам реальных испытаний можно известными методами с определенной точностью количественно оценить аргументы этой функции, то есть оценить уровень знаний испытуемых и уровни трудностей заданий теста. Если в формуле (1) зафиксировать переменную β_j , а θ рассматривать как независимую переменную, то получим

$$P_j = P_j(\theta) = \varphi(\theta - \beta_j) \quad (2)$$

вероятность правильного выполнения испытуемыми с различными уровнями подготовки j -го задания. Примерный график функции (2) представлен на рис. 1 и называется характеристической кривой j -го задания теста.

Таким образом, если уровень знаний испытуемого θ намного больше уровня трудности данного β_j , то вероятность его правильного ответа на это задание $P_j \rightarrow 1$. Аналогично, если θ намного меньше β_j , то $P_j \rightarrow 0$. Если же $\theta = \beta_j$, то $P_j = 0,5$ (см. рис. 1).

Если в формуле (2) зафиксировать переменную θ а β рассматривать как независимую переменную, то получим

$$P_i = P_i(\beta) = \varphi(\theta_i - \beta) \quad (3)$$

вероятность правильного выполнения испытуемым i различных по трудности заданий теста. Примерный график функции (3) представлен на рис. 2 и называется характеристической кривой испытуемого i .

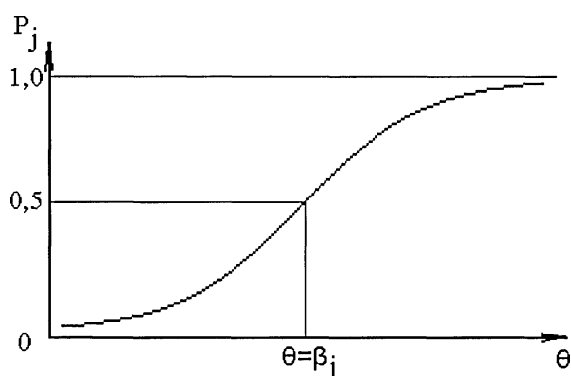


Рис. 1. Характеристическая кривая j -го задания теста

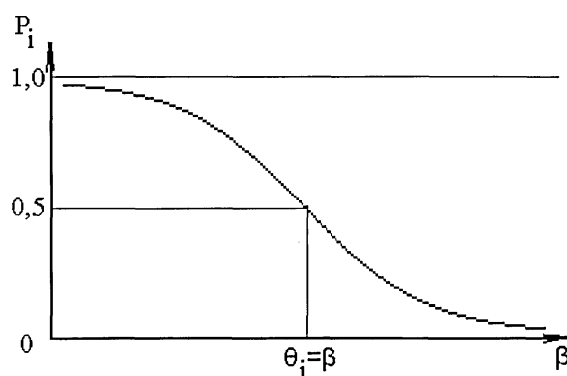


Рис. 2. Характеристическая кривая испытуемого i

Таким образом, если уровень трудности задания β намного больше уровня подготовленности данного испытуемого θ то вероятность его правильного ответа на это задание P_i , стремится к 0.

Библиографический список

1. Михеев, О.В. Математические модели педагогических измерений / О.В. Михеев // Педагогические измерения. – 2004. – № 2. – С. 75–88.
2. Нейман, Ю.М. Введение в теорию моделирования и параметризации педагогических тестов / Ю.М. Нейман, В.А. Хлебников. – М., 2000. – 168 с.
3. Чельшкова, М.Б. Теория и Практика конструирования педагогических тестов / М.Б. Чельшкова. – М.: Логос, 2002. – 432 с.
4. Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (2005). Adapting educational and psychological tests for cross-cultural assessment. Mahwah, NJ: Lawrence Erlbaum. 378 + xii pp., 26 b&w illus.