УДК 378.14 + 158

МОДЕЛИРОВАНИЕ ЭФФЕКТИВНЫХ ТЕСТОВ

© В.В. Зубец

Zubets V.V. Designing effective tests. The paper studies the ways to increase the quality of educational tests on the basis of testing results correlation analysis. In order to create quality educational tests, the author proposes a step-by-step approach to the «ideal» model.

I. ИСТОРИЯ ВОПРОСА

История тестирования насчитывает более ста лет. Первые тесты появились в конце XIX века и призваны были стать новым методом оценки различных свойств личности. С начала XX века начинает формироваться классическая теория тестирования, базировавшаяся на теории измерений и теории ошибок [1–5]. Тесты стали применяться в различных областях человеческой деятельности: медицине, психологии, в том числе и в образовании для контроля знаний. Принципиальное отличие тестирования от традиционных методов контроля знаний в виде оценки знаний обучаемого преподавателем (экспертной оценки) заключается в том, что только для тестирования разработана научно обоснованная система оценки качества самого тестирования, что позволяет перейти от качественной субъективной оценки знаний учащихся к объективным количественным измерениям результатов обучения. Именно поэтому педагогическое тестирование относят к методам объективного контроля (МОК) знаний и навыков. Понятно, что нельзя организовать эффективное управление образовательным процессом, не имея такого метода контроля. Зародившись на Западе, тестирование стало развиваться и в России. К сожалению, в 30-е годы XX века в СССР методы объективного контроля знаний были признаны «буржуазными и вредными» по идеологическим причинам [6]. Это сильно замедлило распространение МОК в отечественном образовании. Между тем, на Западе тестология не стояла на месте, в 70-х годах XX века появилась новая теория тестирования Item Response Theory (IRT) [7–11], основанная на построении математико-статистических моделей поведения обучаемого, согласовании этих моделей с результатами эмпирических измерений с целью выявления латентных (скрытых) качеств обучаемого. В настоящее время МОК нашли широкое применение в образовании развитых стран и доказали свою эффективность.

В последние годы в российской сфере образования наблюдается усиление интереса к педагогическим тестам. Снятие идеологических препятствий привело к росту числа исследований методов объективного контроля знаний. Ситуация изменилась на противоположную; если раньше чиновники запрещали тестирование, то теперь именно министерство образования является одним из главных инициаторов внедрения тестов в образовательный процесс. В целом, достигнуты определенные практические результаты: это и проводимое в течение ряда лет централизованное тестирование, и программа перехода к единому экзамену. На пути внедрения МОК в образование встречается немало трудностей. С одной стороны, это сопротивление противников тестирования, а с другой, – ошибки энтузиастов-реформаторов, видящих в применении тестов панацею от всех бед. Дискуссия сторонников и противников тестирования часто ведется беспредметно. Одни говорят, как хорошо тестирование, другие – как оно ужасно. Между тем, если вести речь о конкретной задаче, все становится на свои места. Каждый метод контроля результатов обучения имеет свою область применения, в пределах которой его использование наиболее эффективно, соответственно, есть области, где его применение неэффективно. Так, даже в США, где тестирование очень широко распространено, применяются и другие методы контроля. Сложные виды творческой деятельности, например, способность специалиста самостоятельно осуществить проект, невозможно однозначно оценить набором тестовых заданий. Для этого значительно эффективнее будет такая форма экспертной оценки, как курсовая, дипломная или диссертационная работа, что и делается на практике. С другой стороны, если необходимо проверить освоение обучаемыми тезауруса специальной области, применение тестов чрезвычайно эффективно, поскольку процедура такой проверки достаточно механическая, всем обучаемым необходимо задавать одни и те же вопросы. Правильное применение тестирования освобождает время преподавателя от рутинной, механической работы (ее берет на себя компьютер), что позволяет заняться сложной экспертной оценкой, непосильной для компьютерных программ.

Существует и другая проблема – качество тестов. Сильной стороной тестирования является то, что тестовые задания поддаются строгой научной оценке, то есть можно количественно оценить качество теста. К сожалению, лишь немногие педагоги владеют методами оценки качества тестов, поэтому на практике нередко используются либо неэффективные задания, либо задания, которые с точки зрения тестологии тестовыми вообще не являются.

II. ПОСТАНОВКА ЗАДАЧИ

В данной работе исследовались способы повышения качества педагогического теста. В качестве исследования были выбраны навыки, которые приобретали

студенты филологического факультета Тамбовского государственного университета имени Г. Р. Державина на занятиях по информатике. Применение теста для контроля результатов обучения в данном случае не только оправдано, но и необходимо. Дело в том, что в течение курса информатики студенты осваивают около пятидесяти различных команд. На контроль результатов обучения (зачет) одного студента отводится 5,5 минут. И это, не считая того, что необходимо проверить не только практические навыки, но и знание теоретического материала. Понятно, что проверить 50 навыков менее чем за 6 минут невозможно. Поэтому преподаватель вынужден либо делать неполный (выборочный) контроль, либо автоматизировать данный процесс с помощью компьютерного теста, что позволит быстро проверить одновременно целую группу студентов.

Процессы ответа на тестовые задания и выполнения команд при экспертном контроле сильно различаются. Насколько сильно они различаются и можно ли это различие преодолеть, необходимо было выяснить. Итак, задачу можно сформулировать так. Идеальной формой контроля навыков является экспертная оценка, которая заключается в том, что преподаватель (эксперт) непосредственно проверяет умение каждым студентом выполнить каждую изученную команду. Валидность такой процедуры близка к 1, поскольку установить, умеет ли студент выполнять ту или иную команду, можно точно. Метод экспертной оценки точен, но чрезвычайно расточителен по времени, так как один эксперт одновременно может проверять умение выполнять одну команду одним студентом. Делается предположение, что можно составить компьютерный контролирующий тест, который будет моделировать процесс экспертной проверки. Тест позволит с большой скоростью проверять одновременно большое количество студентов, что значительно повысит экономическую эффективность контроля. При этом, чем сильнее корреляция результатов тестирования и экспертной оценки, тем модель более адекватна процессу экспертизы.

III. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТА

Эксперимент проводился следующим образом. В течение 3–4 занятий студенты практически осваивали в компьютерном классе работу с файлами и каталогами и приобретали определенные навыки. Для проверки приобретенных на занятиях навыков были выбраны 13 команд операционной системы: 1 – смена панелей, 2 – просмотр файла, 3 – редактирование файла, 4 – вход в каталог, 5 – выход из каталога, 6 – создание каталога, 7 – создание файла, 8 – смена диска, 9 – удаление файла, 10 – определение загруженности диска, 11 – отображение дерева каталогов, 12 – копирование файлов, 13 – перемещение файлов. Был составлен компьютерный тест, включающий по два задания на каждую из команд. Одно из заданий было на выбор правильного варианта (тип а), а другое – на установление правильной последовательности действий (тип в). Сразу же после тестирования проводились экспертные оценки навыков по тем же 13-ти командам. В опросных листах проставлялись оценки по принципу «умеет – не умеет» (1 или 0). Объем выборки составил 100 испытуемых. Тестирование проводилось дважды: перед обучением и спустя 3–4 занятия.

Результат обучения обычно характеризуется изменением меры трудности в результате обучения. Мера трудности определяется как доля неправильных ответов. Чем выше мера трудности, тем ниже уровень обученности. Обработка результатов тестирования показала, что если перед обучением средняя мера трудности равна 0,8, то после занятий она понижается до 0,3, то есть почти в три раза. Таким образом, тестирование четко фиксирует результат обучения. Нередко на практике преподаватели этим и довольствуются: есть зависимость между обучением и мерой трудности и хорошо.

Согласно классической теории тестов, важнейшими характеристиками тестов являются надежность и валидность. Надежность характеризует повторяемость результатов, а валидность – уверенность в том, что тест измеряет именно ту характеристику, которую планировалось измерять при его создании.

Оценка надежности по формуле Кудера–Ричардсона дает значение 0,7, что является вполне приемлемым:

$$R = k/(k-1)\cdot(1 - \Sigma p_j g_j / S^2),$$

где R – коэффициент надежности, k – число заданий теста, $\Sigma p_j g_j$ – сумма дисперсий заданий теста, S^2 – общая дисперсия баллов испытуемых по всему тесту.

Теперь о валидности. Понятно, что умение практически выполнить команду и описать этот процесс словами не одно и то же. Поэтому оценка валидности для теста на проверку приобретенных навыков очень важна. Для определения валидности можно сравнить результаты тестирования и экспертной оценки. Если принять валидность экспертной оценки за 1, то коэффициент корреляции результатов тестирования с экспертной оценкой будет характеризовать валидность теста. Определение коэффициента корреляции между мерами трудности заданий теста и экспертных оценок дает значение 0,6, что, казалось бы, говорит о достаточно высокой валидности. Однако специалисты рекомендуют оценивать валидность как коэффициент корреляции между тестовыми баллами студентов и их экспертными оценками [6]. Эта корреляционная связь должна быть сильной, так как тестовый балл студента, показавшего хорошую подготовку эксперту, должен быть выше, чем у студента с плохой подготовкой. Однако расчет коэффициента корреляции дает очень низкое значение – 0,14.

Приведенные характеристики теста показывают, что он неплохо измеряет уровень обученности студентов, обладает удовлетворительным значением надежности, но валидность его низка.

Таблица 1

Характеристики тестовых заданий типа а

Номер команды	1	2	3	4	5	6	7	8	9	10	11	12	13
Мера трудности	0,1	0,3	0,1	0,2	0,2	0,2	0,1	0,1	0,2	0,5	0,3	0,3	0,3
Валидность	0,2	0,3	0,3	0,3	0,4	0,4	0,2	0	0,2	0,1	0,2	0,1	0,2

Таблица 2

Характеристики тестовых заданий типа в

Номер команды	2	3	4	5	6	7	8	9	10	11	12
Мера трудности	0,2	0,1	0,1	0,2	0,1	0,2	0,2	0,1	0,8	0,6	0,3
Валидность	0,3	0,2	0,4	0,2	0,4	0,2	0,2	0,2	0,2	0,3	0,4

Таблица 3

Коэффициенты корреляции заданий «улучшенного» теста

Номер команды	2 (а)	3 (а)	4 (а)	5 (а)	6 (а)	7 (в)	11 (в)	12 (в)
Валидность	0,5	0,4	0,5	0,5	0,6	0,6	0,4	0,5

Рассмотрим внимательнее характеристики отдельных заданий. Результаты эксперимента можно представить в виде матрицы, столбцы которой представляют собой результаты тестирования по конкретным заданиям, а строки – результаты по отдельным студентам. Структура матрицы результатов следующая: столбцы 1–25 – задания тестов типа а и в, столбец 26 – сумма баллов тестовых заданий, столбцы 27–40 – результаты экспертных оценок, столбец 41 – сумма баллов экспертной оценки. Следует отметить, что заданий типа в – 11, а не 13.

В таблицах 1–2 приведены основные характеристики тестовых заданий: мера трудности и валидность. Валидность задания определялась как коэффициент корреляции между соответствующим столбцом матрицы результатов и столбцом суммарной экспертной оценки (41-й столбец).

Анализ результатов показывает, что тест очень неоднороден, характеристики заданий различаются очень сильно, так, мера трудности колеблется от 0,1 до 0,8; значения валидности отдельных заданий низки. Разница между характеристиками заданий различных типов невелика.

Следует отметить, что мера трудности большинства заданий оказалась явно заниженной: 0,1–0,3, тогда как тестологи считают оптимальным значение 0,5 [12, 13]. Из анализа данных можно предположить, что неоднородность характеристик теста связана с неоднородностью заданий. Поэтому когда усреднение производится по большой выборке (100 студентов), то характеристики более качественных заданий теста выявляются, что проявилось в достаточно высоком коэффициенте корреляции меду мерой трудности заданий и экспертными оценками. Если же усреднение происходит по небольшому количеству заданий (25) при расчете валидности, то эти характеристики теряются в шуме от «плохих» заданий.

Чтобы проверить эту гипотезу, была составлена таблица измерений, из которой были убраны самые «плохие» задания, то есть задания с минимальными значениями валидности и меры трудности. После этого были вновь рассчитаны коэффициенты корреляции заданий. Результаты представлены в таблице 3.

Видно, что валидность заданий значительно выросла. Валидность всего теста составила 0,6 для «улучшенного теста» против 0,14 для исходного. Таким образом, валидность «улучшенного» теста значительно выше, чем у исходного, то есть «улучшенный» тест лучше моделирует экспертную проверку навыков.

IV. ВЫВОДЫ

На основе проведенных исследований можно сделать следующие выводы.

1. Возможна разработка педагогических тестов, достаточно хорошо моделирующих процесс приобретения навыков работы на компьютере испытуемыми.
2. Необходимо тщательно проверять характеристики педагогического теста, в частности, полезно рассмотреть матрицу коэффициентов корреляции между тестовыми заданиями, суммарными баллами испытуемых и экспертными оценками. Анализ полученных данных позволит исключить «плохие» задания.
3. Для того чтобы создать качественный тест, необходимо заменить «плохие» задания на более качественные. Иногда для этого необходимо изменить форму задания [14–16]. (В частности, в исследованном тесте отмечалась заниженная мера трудности некоторых заданий. Ее можно повысить, увеличив количество предлагаемых вариантов ответа с 3-х до 5-ти.) После обновления теста следует вновь провести изучение его характеристик.
4. Следует уделять серьезное внимание обучению испытуемых методике тестирования до проведения испытания. Дополнительные исследования показали, что при ответе возникает много технических ошибок. Так, при повторном тестировании, которое проводилось сразу же после первого, результаты получались выше.

Таким образом, для создания качественных педагогических тестов, автор предлагает метод последовательного приближения к «идеальной» модели. Вначале создается тест на основе априорных знаний, затем проводится эксперимент по тестированию и экспертным оценкам, статистический анализ результатов эксперимента, «улучшение» заданий, новый эксперимент и так далее.

Повышение качества педагогических тестов, в конечном счете, должно приводить к повышению качества самого процесса обучения.

ЛИТЕРАТУРА

1. *Cattel J. McK.* Mental Tests and Measurement // Mind. 1890. V. 15. P. 373-380.
2. *Cattel J. McK., Farrand L.* Physical and Mental Measurements of the Students of Columbia University // Psychological Review. 1896. V. 3. № 6.
3. *Kelley T.L.* Interpretation of Educational Measurements. N. Y.: World Books Co, 1927. 363 p.
4. *Kuder G.F., Richardson N.W.* The Theory of Estimation of Test Reliability // Psychometrika. 1937. V. 2. P. 151-160.

5. *Lawley D.N.* On Problems Connected with Item Selection and Test Construction // Proceedings of the Royal Society of Edinburg. Section A Mathematical and Physical Sciences. 1943. V. 43. LXL. Part III. P. 273-287.
6. *Аванесов В.С.* Научные основы тестового контроля знаний. М.: Исследовательский центр, 1994. 135 с.
7. *Bollen K.A.* Structural Education with Latent Variables. N. Y.: Wiley & Sons, 1989. 514 p.
8. *Lord F.M.* Application of Item Response Theory to Practical Testing Problems. Hillesdale N. J.: Lawrence Erlbaum Ass., Publ. 1980. 266 p.
9. *Rash G.* On General Laws and the Meaning of Measurement in Psychology. Berkley: Univ. of California Press, 1961.
10. *Rash G.* On Specific Objectivity: An Attempt of Formalizing the Regrets for Generality and Validity of Scientific Statements // Danish Yearbook of Philosophy. Copenhagen: Munkagaard, 1977. V. 14. P. 58-94.
11. *Rash G.* Probabiliatic Models for Some Intelligence and Attainment Tests / With a Foreword and Afterword by B.D. Wright. Chicago & London: The Univ. of Chicago Press, 1980. 199 p.
12. *Аванесов В.С.* Композиция тестовых заданий. М.: Адепт, 1998. 217 с.
13. *Клайн П.* Справочное руководство по конструированию тестов. Киев, 1994.
14. *Зубец В.В., Ильин А.А.* Изучение некоторых свойств компьютерных тестов // Актуальные проблемы информатики и информационных технологий: Материалы IV-ой Тамбов. межвузов. науч. конф. Сентябрь 2000 г. Тамбов: ТГУ, 2000. С. 22-23.
15. Организация тестового контроля: Учебн.-методич. пособие / Авт.-сост. Н.В. Кузьмина., М.С. Чванова, В.В. Зубец. Тамбов: Изд-во ТГУ, 1988. 42 с.
16. *Зубец В.В.* Применение корреляционного анализа для повышения качества педагогических тестов // XI Междунар. конф.-выставка «Информационные технологии в образовании (ИТО – 2001)»: Сб. тр. участников конф. Ч. V. М.: МИФИ, 2001. С. 32-33.

Поступила в редакцию 10 ноября 2001 г.