А.А. Малыгин

ТЕОРИЯ ПЕДАГОГИЧЕСКИХ ИЗМЕРЕНИЙ КАК НАУЧНАЯ ОСНОВА ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ КОНТРОЛЬНО–ОЦЕНОЧНЫХ ПРОЦЕССОВ В ОБРАЗОВАНИИ

Ивановский государственный химико-технологический университет
E-mail: a_malygin@mail.ru

В статье рассматриваются вопросы, связанные с повышением эффективности контрольно-оценочных процессов в образовании. Подчеркивается ключевая роль объективного оценивания, научной основой которого является современная теория тестов. Приводятся обоснования в пользу выбора компьютерного адаптивного тестирования.

Ключевые слова: эффективность контроля в образовании, современная теория тестов, педагогические измерения, компьютерное адаптивное тестирование.

Наука начинается с измерения…
Д.И. Менделеев

Современная ситуации в образовании характеризуется переходом на новые образовательные стандарты, предписывающие необходимость компетентностной ориентации не только образовательного процесса, его содержания и технологий реализации, но и перестройку контрольно-оценочных процессов, технологий и средств качества подготовки обучающихся. Это должно отражаться на всех стадиях образовательного процесса: от входной аттестации, через все виды промежуточной аттестации до итоговой аттестации на соответствие требованиям федеральных государственных образовательных стандартов (ФГОС) [1]. При этом в центре внимания находятся вопросы качества образования и его оценка, которая предполагает целенаправленное рассмотрение всех компонентов образовательных систем, где ключевая роль отводится объективному оцениванию, основанному на теории педагогических измерений и обеспечивающему научный базис для анализа, функционирования, развития, прогнозирования и совершенствования систем управления качеством образования. Ядро научного обоснования образует методология педагогических измерений, под которой следует понимать систему основных положений, принципов, форм и методов, нацеленных на достижение максимально возможной эффективности измерений за счет научной организации практики в сфере контрольно-оценочных процессов.

Информация, накапливаемая в рамках научно-обоснованной контрольно-оценочной деятельности, должна обладать репрезентативностью, обоснованностью, прогностичностью и высокой объективностью, а также сочетать количественные и качественные методы, средства и технологии, которые нашли свое отражение в бипарадигмальной методологии педагогических измерений, получившей широкое развитие в образовании с начала XXI века [2].

В теории педагогическое измерение понимается как процесс установления соответствия между оцениваемыми характеристиками и точками шкалы, в которой отношение между различными оценками выражается свойствами числового ряда. Процесс педагогических измерений включает в себя: объект измерения (одну или несколько латентных характеристик), измерительные процедуры, измерительные инструменты (тест и шкалу для фиксации оценок измеряемого объекта), анализ и интерпретацию результатов измерения. Перечисленные компоненты процесса измерения имеют свои аналоги в традиционном контроле, однако там эти процедуры больше носят интуитивный характер. При тестировании каждая компонента должна проходить процесс научного обоснования

качества, особенно важно, если речь идет об итоговом контроле, результаты которого служат для принятия управленческих решений. В этом случае объектом измерения становятся знания, умения, навыки и компетенции, часто называемые обобщающим термином «учебные достижения», структуру и уровень которых сравнивают с требованиями ФГОС.

При педагогических измерениях возникает как минимум три ключевых вопроса, продиктованных требованиями научности и эффективности информации: что такое объективное измерение? Что является переменной измерения и измеряется ли эта переменная при выполнении теста? Как разработать иснтрумент, обладающий планируемыми характеристиками и обеспечивающий оценивание с заданной точностью? Конечно, дать развернутые ответы на поставленные вопросы в одной статье просто не удастся, но попытаемся ввести некоторые подходы и обосновать использование математического аппарата современной теории тестов для повышения эффективности контрольно-оценочных процессов в образовании.

Как известно, объективность – это недостижимое свойство в любых измерениях, в том числе и педагогических, поскольку на результаты измерений всегда оказывается множество систематических и случайных ошибок. В связи с этим, когда говорят об объективных результатах тестирования, то чаще подразумевают процедурную объективность, предполагающую исключение педагога из контрольно-оценочного процесса и применение компьютерных технологий. Поэтому правильнее следует говорить о высокой или низкой объективности, но при этом высокая объективность не является неотъемлемым свойством тестовых оценок.

Если ввести в рассмотрение истинный балл испытуемого, трактуя его как свободный от ошибки результата измерения, то наблюдаемый балл будет отличаться от истинного на величину ошибки измерения. В случае, когда ошибка измерения превышает допустимые пределы заданной точности, не приходится говорить о достижении высокой объективности. Возможность оценивания ошибки измерения является ключевым критерием, позволяющим отделить оценки, которые можно считать результатами измерения, от результатов традиционного контроля. Вполне очевидно, что традиционные средства контроля не обеспечивают никаких данных о точности оценок достижений учащихся в отличие от тестов, позволяющих установить величину ошибки измерения и оценить тем самым надежность полученных оценок. Поэтому при педагогических измерениях принято говорить о надежности (или ненадежности), определяемой как степень устойчивости (повторяемости) и точности результатов измерения.

Углубленная трактовка термина «объективность измерений» лежит в области современной теории конструирования тестов (Items Response Theory - IRT). Благодаря свойствам симметричности и независимости измеряемых латентных переменных, в роли которых выступают параметры подготовленности испытуемых и трудность заданий теста, при использовании IRT снижается ошибка измерения и повышается объективность получаемых оценок параметров. Независимость оценок латентных параметров, реализуемая за счет специальных свойств математических моделей IRT, приводит к инвариантности оценок параметра испытуемых относительно трудности заданий теста. Поскольку в IRT существует дифференцированная ошибка измерения, объективность проявляется по-разному для выборки испытуемых при выполнении одного и того же теста. В центре распределения баллов испытуемых, где ошибка меньше, будут получены более достоверные результаты. На краях распределения ошибка измерения будет возрастать, и соответственно, будет снижаться объективность измерения.

Для повышения точности оценок и тем самым эффективности контрольно-оценочных процедур, следует отдать предпочтение измерениям, основанным на применении IRT. На практике достижение высокой объективности (или надежности) достаточно сложный процесс, поскольку требует привлечение обширного математико-статистического аппарата IRT, проведение тщательного анализа и коррекции распределения эмпирических данных тестирования и специальных программных продуктов обработки и интерпретации результатов выполнения тестов.

На практике у преподавателя логично возникнет вопрос о том, какого рода объективность и какие измерители следует использовать при оценивании учебных достижений. Как отмечалось выше, измерение может проводиться на разных уровнях – количественном и качественном, во многом сходных по целям осуществления, но имеющие существенные отличия по проявлению свойств оцениваемых латентных характеристик. В соответствии со свойствами для качественного (неметрического) и количественного (метрического) уровня измерений существенно отличаются способы измерения, формы представления результатов, группы допустимых операций с различными статистическими величинами. Отличие уровня измерений проявляется и в выборе измерителей. Для качественных измерений разрабатывают компетентностные тесты, портфолио, кейс-измерители и анкеты, проводят

интервьюирование, собеседования и традиционные экзамены. К качественному уровню измерений принято относить также результаты тестирования, если в них включены задания со свободно конструируемыми ответами, так как для проверки ответов привлекаются эксперты, которые вносят субъективный фактор. В количественных измерениях используются стандартизированные тесты с выбором ответов, автоматизированные формы проверки и обработка данных при шкалировании на основе IRT.

Для любой, в том числе и внутривузовской, контрольно-оценочной системы в образовании можно выделить входные тесты, тесты для текущей работы и итоговые тесты. Для каждого из перечисленных видов тестов выстраиваются определенные стандарты качества, соблюдение которых в процессе разработки и применения тестов позволяет достичь запланированной точности измерения и дает возможность судить о достигнутой объективности результатов тестирования.

Особо следует отметить важность соблюдения требований к качеству контрольно-оценочных процедур при аттестации, где высока степень ответственности принимаемых управленческих решений и требуется высокое качество измерений. Система государственной аккредитации, единый государственный экзамен и другие массовые процедуры тестирования, выходящие по сфере интерпретации результатов на внешнего потребителя регионального и федерального уровня, требует обращения при измерениях к более высокому уровню объективности, предполагающему использование IRT. Но для применения IRT необходимо понимать, что формальное обращение к моделям не обеспечит автоматическую минимизацию ошибок измерения. Инвариантность оценок латентных параметров испытуемых и заданий, лежащая в основе введения объективности, реализуется далеко не во всех моделях IRT.

В современной теории тестов IRT латентный параметр испытуемого (истинный балл) трактуется как некоторая переменная, начальная оценка которого получается непосредственно из эмпирических данных тестирования. Переменный характер измеряемой величины указывает на возможность последовательного приближения к объективированным оценкам параметра с помощью определенных итерационных методов. В педагогических тестах, являющихся основным инструментом измерений, в качестве латентной переменной выступает уровень подготовленности, который в IRT обозначается символом θ, а латентный параметр трудности задания – β, которые распределены по нормальному закону. Датский математик G. Rash, занимающийся вопросами оптимального соотношения между параметрами θ и β, предложил форму связи параметров в виде разности $\theta–\beta$. Введение разности для оценок трудности заданий и уровня подготовленности предполагает существование единой интервальной шкалы с единицей измерения, получившей в работах зарубежных исследователей название «логит» [5].

Выбор математической модели, описывающей взаимосвязь между эмпирическими результатами тестирования и значениями латентных параметров θ и β является центральным в IRT. В соответствии с основным предположением существует некоторая взаимосвязь эмпирическими результатами тестирования и значениями латентных параметров θ и β, которая может быть выражена через математическую модель. Исследования A. Birnbaurm, F.M. Lord, G. Rash по анализу линии регрессии наблюдаемых результатов выполнения теста на латентную переменную θ привели к выводу о нелинейном характере связи между наблюдаемыми и истинными баллами.

Относительная инвариантность значений латентных переменных от конкретного тестирования, определенная устойчивость частот появлений их значений послужили основанием для использования понятия вероятности события как меры возможности его появления. В качестве такого события исследователи выбрали правильный ответ i-го испытуемого на j-е задание теста. Можно рассматривать условную вероятность правильного выполнения i-м испытуемым с уровнем подготовленности θ_i различных по трудности заданий теста, считая θ_i - параметром i-го испытуемого, а β - независимой переменной. Тогда условная вероятность P_i будет функцией латентной переменной β:

$$P_i\{x_{ij} = 1 \mid \theta_i\} = f(\theta_i - \beta), i = 1,2,..., N \qquad (1)$$

Аналогично вводится условная вероятность правильного выполнения j-го задания трудностью β_j различными испытуемыми тестируемой группы. В данном случае независимой переменной является θ, а β_j – параметр, определяющий трудность j-го задания теста. Тогда

$$P_j\{x_{ij} = 1 \mid \beta_j\} = F(\theta - \beta_j),\ j = 1,2,..., n, \qquad (2)$$

где: $x_{ij} = \begin{cases} 1, \textit{если ответ } i-\textit{го испытуемого} \\ \textit{на } j-\textit{ое задание верный}; \\ 0, \textit{если ответ } i-\textit{го испытуемого} \\ \textit{на } j-\textit{ое задание неверный} \end{cases}$

N – число испытуемых, n – количество заданий в тесте.

В теории IRT функции (1) и (2) обозначаются как $P_i=f(\beta)$ и $P_j=F(\theta)$ соответственно и называются Item response functions (IRF). Графиком первой функции является убывающая индивидуальная кривая испытуемого (рис. 1), а вторая возрастающая функция – это характеристическая кривая задания (рис. 2).

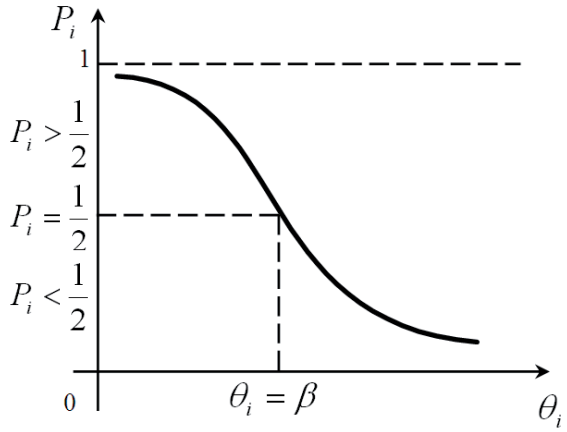

Рис. 1. График функции $P_i=f(\beta)$ (индивидуальная кривая испытуемого)

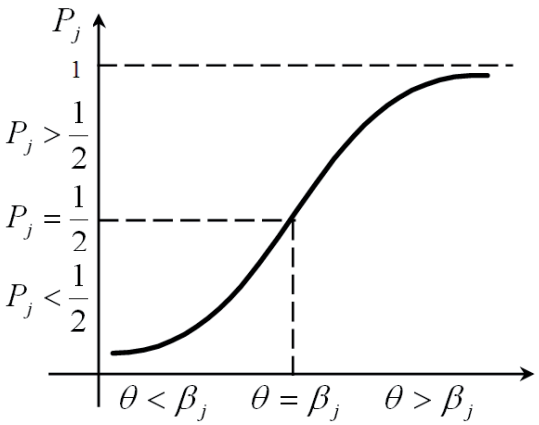

Рис. 2. График функции $P_j=F(\theta)$ (характеристическая кривая задания)

Число параметров, входящих в аналитическое задание функций, является основанием для подразделения семейства IRF на классы. Среди логистических функций различают несколько наиболее удобных для практического использования, к числу которых принадлежат однопараметрическая модель G. Rasch, двух и трехпараметрические модели A. Birnbaurm [5].

Сообразно различным целям в образовательном процессе выбираются различные параметрические модели. Наиболее эффективными для целей управления качеством образования представляются динамические модели, строящиеся на основе модели G. Rasch с помощью введения дополнительного параметра времени. Общий вид динамической модели, предложенной Fischer, представлен ниже

$$P_j(x_{ij}=1)=\frac{e^{(\theta-\delta)-\beta_j}}{1+e^{(\theta-\delta)-\beta_j}} \qquad (3)$$

где $P_j(x_{ij}=1)$ – вероятность правильного ответа на j-е задание; θ – параметр подготовленности испытуемых, оценка которого указывает на уровень учебных достижений; δ – параметр, учитывающий фактор времени и позволяющий оценить прирост учебных достижений; β_j – параметр трудности задания.

С целью индивидуализации и гуманизации обучения, которое может протекать также в дистанционном режиме для различных участников образовательного процесса, предпочтительнее всего переходить от стандартного тестирования с конечным набором заданий к компьютерному адаптивному тестированию (КАТ) [6, 7]. Часто и ошибочно под адаптивностью понимают пошаговый отбор заданий, основанный на дихотомической оценке результатов их выполнения (выполнил – не выполнил). В действительности такой подход далек от научного понимания адаптивного тестирования. Дело в том, что достичь оптимизации, а тем самым и эффективности контроля при адаптивном тестировании, позволяют именно такие задания, которые не оцениваются дихотомически и имеют устойчивые параметры их трудности. Поэтому адаптивное тестирование предполагает использование вероятностных моделей IRT. В общем случае адаптивное тестирование следует определить как совокупность автоматизированных процессов генерации, предъявления и оценки результатов выполнения адаптивных тестов, обеспечивающую прирост эффективности измерений по сравнению с традиционным тестированием за счет оптимизации подбора характеристик заданий, их количества, последовательности и времени предъявления применительно к уровню подготовленности испытуемых.

Эффективность контрольно-оценочных процедур повышается при использовании многошаговых стратегий отбора и предъявления заданий, основанных на полной контекстной зависимости, в которых каждый очередной шаг совершается только после оценки результатов выполнения предыдущего шага. После выполнения испытуемым очередного задания принимается разветвляющееся решение о подборе трудности следующего задания в зависимости от того,

верным или неверным был предыдущий ответ. Реализация таких алгоритмов возможна лишь при наличии банка калиброванных заданий с устойчивыми оценками их параметров, использование специальных компьютерных программ для индивидуализации алгоритмов подбора заданий и применение параметрических моделей IRT.

В целом к числу важных преимуществ в пользу компьютерного адаптивного тестирования в контрольно-оценочном процессе следует отнести следующие:

- высокую эффективность, обеспечиваемую за счет минимизации числа заданий, времени тестирования в условиях, когда ошибка измерения не выше (либо равна, либо ниже) той, которая получается в аналогичных по содержанию традиционных тестах;
- высокий уровень секретности, практически исключающий возможность списывания, подсказок и других нежелательных эффектов в процессе выполнения адаптивных тестов;
- индивидуализацию темпа выполнения теста, обеспечиваемую адаптивными алгоритмами и соответствующим программным обеспечением, с помощью которых подбор очередного по трудности задания происходит только после выполнения предыдущего задания адаптивного теста;
- повышение уровня мотивации к тестированию у наиболее слабых учащихся за счет исключения предъявления излишне трудных заданий, способствующих росту фактора тревожности и чувства страха;
- сообщение результата в интервальной шкале тестовых баллов каждому тестируемому незамедлительно, сразу после окончания его работы над индивидуально подобранным набором заданий в адаптивном тесте;
- исключение временных, организационных и финансовых затрат на стандартизацию для установления норм тестов в силу отсутствия традиционных тестов фиксированной длины.

В исследованиях, которые проводились при участии автора статьи, были апробированы различные виды тестирования с позиции их эффективности, которые привели к выбору компьютерного адаптивного тестирования с использованием многошаговых варьирующих стратегий. Наиболее их важное преимущество заключается в возможности оперативного реагирования на результаты выполнения тестовых заданий путем переоценки уровня подготовленности учащегося после выполнения каждого очередного задания теста. Данное обстоятельство позволяет соотнести качество структуры знаний учащихся с длиной индивидуальной траектории в процессе генерации адаптивного теста: чем меньше крутизна индивидуальной кривой тестируемого (рис. 1), тем длиннее его индивидуальная траектория выполнения заданий теста. Следовательно, можно определить оптимальное число заданий, необходимых для контроля каждого студента, что доказывает эффективность использования адаптивного тестирования.

Разумеется, что корректный выбор моделей измерений не решает всех проблем, связанных с повышением эффективности контроля и оценки качества учебных достижений. Помимо выбора моделей и разработки измерителей, обеспечивающих запланированную точность оценок, необходимо также думать о шкалировании результатов тестирования, обеспечивающем корректность и удобство интерпретации баллов испытуемых и эффективность управленческих решений. В этой связи, наиболее перспективными представляются многоуровневые шкалы, которые можно будет сделать интервальными лишь после создания всего комплекса тестов, включая и междисциплинарные тесты.

Список использованной литературы

1. Ефремова Н.Ф. Подходы к оцениванию компетенций в образовании: учеб. пособие. М.: ИЦ ПКПС, 2010.
2. Звонников В.И. Измерения и качество образования: монография. М.: Логос, 2006.
3. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов. М.: Логос, 2010.
4. Челышкова М.Б. Адаптивное тестирование в образовании (теория, методология, технология). М.: ИЦ ПКПС, 2001.
5. Lord F.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale N.-J.: Lawrence Erlbaum Ass., Publ., 1980.
6. Wainer H. Computerized Adaptive Testing: A Primer. Springer, 2000.
7. Weiss D.J. (Ed.). New Horizons in Testing: Latent Trait Test Theory and Computerised Adaptive Testing. N–Y.: Academic Press, 1983.

Статья получена 18.10.2010
Принята в печать 22.10.2010