УДК 004.891.3

# АНАЛИЗ СУЩЕСТВУЮЩИХ МЕТОДОВ ОЦЕНКИ КАЧЕСТВА ТЕСТОВЫХ МАТЕРИАЛОВ

***С.В. Окладникова***

*В статье рассматриваются вопросы, связанные с оценкой качества тестовых материалов при использовании экспериментально-статистических методов и методов экспертного оценивания. Проводится сравнительный анализ достоинств и недостатков указанных методов, а также описываются особенности их практического использования.*

Ключевым моментом диагностики знаний учащихся с применением методов тестирования является анализ качества тестовых материалов, необходимый для адекватной оценки уровня знаний испытуемых. Оценка качества тестовых материалов в настоящее время выполняется на основе экспериментально-статистических методов и методов экспертного оценивания.

При использовании экспериментально-статистических методов оценка качества тестовых материалов заключается в эмпирическом исследовании статистических характеристик тестов и тестовых заданий, по результатам которого выявляются и исключаются из теста экстремальные задания (т.е. задания, на которые не ответил ни один испытуемый или ответили все испытуемые); проводится анализ совместимости тестовых заданий; оценивается равномерность распределения заданий по трудности; оценивается соответствие уровня трудности теста подготовленности испытуемых.

Расчеты выполняются в соответствии с положениями классической и современной систем тестирования. Оба метода базируются на статистической обработке первичного балла, т.е. балла, набранного испытуемыми в результате предварительного тестирования.

Классический подход к интерпретации результатов подробно освещен в литературе по вопросам статистической обработки данных в педагогике и психологии и по вопросам педагогических измерений[1].

Процесс обработки результатов тестирования включает 10 этапов, на которых определяют: дисперсию каждого задания теста; стандартную ошибку измерения; среднее квадратичное отклонение результатов от средней величины; доверительный интервал для истинных компонентов измерения; трудность каждого тестового задания; степень корреляции тестовых заданий между собой. На основе полученных значений оценивают надежность и валидность тестовых материалов.

Основным достоинством классического метода выступает простота и прозрачность модели. Однако расчет всех вышеперечисленных показателей зависит от параметров сформированного теста, а также условий проведения тестирования.

Методика оценки качества тестового вопроса (или теста в целом) согласно классической теории тестирования предполагает формирование матрицы результатов, исходя из условия, что всем студентам в процессе проведения тестирования предлагается один и тот же сформированный набор заданий. Тест формируется из ограниченного набора тестовых заданий $\boldsymbol{m}$, которые испытуемый должен пройти в течение заданного времени $\boldsymbol{t}$. Тестовые задания могут предъявляться испытуемому либо в заданной последовательности, либо в случайном порядке.

В зависимости от целей (промежуточная или итоговая аттестация) время тестирования целесообразно устанавливать в диапазоне 10–60 мин. Ограничение времени в 60 мин в процессе обучения связано, в первую очередь, с временными рамками проведения занятия (70–90 мин), во-вторых, с утомляемостью тестирующихся, потерей внимания при длительной работе за компьютером и, как следствие, ухудшением результатов тестирования.

Лимит времени проведения тестирования накладывает и ограничение на длину теста. Например, при ограничении времени 30 мин количество вопросов (среднего уровня сложности) в тесте должно составлять 20–30. Это объясняется особенностями технологии компьютерного тестирования: студенту необходимо дополнительное время для прочтения

вопроса с экрана монитора, обдумывания вопроса и выполнения действий по выбору (или вводу с клавиатуры) предполагаемых вариантов ответов[2].

Использование небольшого набора вопросов при проведении повторного тестирования в тех же группах (для сбора статистической информации) ведет к запоминанию вопросов и их списыванию, что, в свою очередь, приведет к завышению оценки за тест.

Поэтому было бы разумнее генерировать для каждого студента индивидуальный набор тестовых вопросов из общей базы вопросов по данной тематике (размер которой может быть ограничен техническими параметрами сервера или программным обеспечением системы тестирования). Например, в результате проведения тестирования студенту предлагается ответить на $m$ тестовых вопросов из множества $n$ вопросов в базе по выбранной теме. В этом случае при формировании индивидуального набора тестовых заданий (случайная выборка, предъявление вопросов по уровню сложности и т.д.) нарушается условие формирования матрицы, по которой обрабатываются результаты предварительного тестирования согласно классической теории тестов, а именно: существуют тестовые вопросы, которые не были сформированы ни в один из тестовых наборов или не имеют достаточных данных для статистической обработки. С другой стороны, при использовании алгоритма случайная выборка (или при повторных тестированиях по теме) некоторые тестовые вопросы попадались одному и тому же студенту более чем один раз, а другому – ни разу. Следовательно, возникает проблема с расчетом индивидуальных баллов студентов и правильных ответов по тесту. Было бы неправильным проводить оценку качества вопроса (согласно классической теории тестирования), который некоторым студентам не попадался. Определение уровня подготовленности учащихся выполняется только по отношению к данному тесту (т.е. по конкретной теме).

Таким образом, в качестве основных недостатков классического метода можно выделить: 1) зависимость оценок от уровня трудности заданий и подготовленности испытуемых[3]; 2) необходимость проведения тщательной калибровки теста, что в реальных условиях организации тестирования не всегда возможно; 3) невозможность сопоставления результатов, полученных по разным тестам, и нанесение их на единую шкалу.

В ***современной теории тестирования*** основное предположение выражается формулой[4]:

$$p_{ij} = f(\theta_i - \beta_j) \qquad (1)$$

где $p_{ij}$ – вероятность того, что $i^{\text{ый}}$ тестирующийся выполнит $j^{\text{ое}}$ задание, $\theta_i$ – латентный (т.е. скрытый от непосредственного наблюдения) параметр способности $i^{\text{го}}$ человека, $\beta_j$ – латентный параметр трудности $j^{\text{го}}$ задания.

Вид функции $f$ различен для разных моделей. Оценка параметров $\theta$ и $\beta$ может проводиться с разной степенью точности – по приближенным формулам или численными методами с заданной точностью. Оценка латентных параметров проводится как по приближенным формулам, так и численными методами, например, методом наибольшего правдоподобия.

Все методы обработки в современной теории тестирования основываются на том, что результат решения отдельного задания испытуемым зависит только от уровня сложности задания и уровня подготовленности испытуемого. Все остальные причины, которые могут иметь влияние на результат выполнения задания (например, физическое и эмоциональное состояние испытуемого, наличие внутренней мотивации выполнения теста и др.), считаются вторичными. Их влияние включается в ошибку измерения, которую можно рассчитать, используя статистические методы.

Уровень сложности задания $\beta_j$ и уровень подготовленности испытуемого $\theta_i$ рассматриваются как переменные (тогда как в классической теории – это некоторые постоянные величины), которые влияют на результат выполнения задания испытуемым и измеряются по единой шкале, называемой шкалой логитов. Такой подход позволяет задать математическую модель, описывающую вероятность $p_{ij}$ верного выполнения $j^{\text{го}}$ задания $i^{\text{м}}$ испытуемым. Значение $p_{ij}$ зависит только от разности $\theta_i$ и $\beta_j$. Функция (1), описывающая зависимость $p_{ij}$ от $\theta_i$ и $\beta_j$, называется логистической, при этом различают несколько видов логистических функций[5]:

- однопараметрическая модель Г. Раша

$$p = \frac{1}{1+e^{-(\theta-\beta)}}; \quad (2)$$

функция зависит от одного параметра – разности $\theta_i$ и $\beta_j$;

- двухпараметрическая модель А. Бирнбаума

$$p = \frac{1}{1+e^{-d(\theta-\beta)}}, \quad (3)$$

где $d$ – дополнительный параметр тестового задания, характеризующий дифференцирующую способность задания;

- трехпараметрическая модель А. Бирнбаума

$$p = c + (1-c)\frac{1}{1+e^{-d(\theta-\beta)}}, \quad (4)$$

где $c$ – еще один дополнительный параметр, характеризующий правильность ответа, если ответ угадан.

Процесс анализа качества тестовых материалов согласно однопараметрической модели Г. Раша заключается в оценке параметров $\boldsymbol{\theta}$ и $\boldsymbol{\beta}$ и нанесении их в одну интервальную шкалу по формулам (4) и (5):

$$\theta_i = \overline{\beta} + \ln\frac{p_i}{q_i}\sqrt{1+\frac{W^2}{2.89}}, i = 1,2,...,N, \quad (4)$$

где $\boldsymbol{N}$ – число испытуемых; $\boldsymbol{p_i}$ – доля правильных ответов $\underline{\boldsymbol{i^{го}}}$ испытуемого на все задания теста; $\boldsymbol{q_i}$ – соответственно доля неправильных ответов; $\overline{\beta}$ – среднее значение логитов трудности заданий теста; $\boldsymbol{W}$ – стандартное отклонение распределение начальных значений параметра β;

$$\beta_j = \overline{\theta} + \ln\frac{q_j}{p_j}\sqrt{\frac{V^2}{2.89}}, j = 1,2,...,n, \quad (5)$$

где $\boldsymbol{n}$ – число заданий; $\boldsymbol{p_j}$ – доля правильных ответов всех испытуемых группы на $\underline{\boldsymbol{j^{ое}}}$ задания теста; $\boldsymbol{q_j}$ – доля неправильных ответов; $\overline{\theta}$ – среднее значение логитов уровней заданий; $\boldsymbol{V}$ – стандартное отклонение распределения начальных значений параметра $\boldsymbol{\theta}$.

На основе полученных данных строятся характеристические кривые заданий теста и испытуемых в предположении нормальности распределений эмпирических данных тестирования, как по множеству испытуемых, так и по множеству заданий. Также считают нормально распределенными и значения латентных переменных. В идеале характеристические кривые должны заполнять равномерно весь интервал шкалы логитов. Данное требование обеспечивается путем устранения из теста заданий, не удовлетворяющих поставленному условию. Однако при изменении различных значений $\boldsymbol{\theta}$ одно и то же задание может оказаться как эффективным, так и неэффективным. Поэтому необходимо более углубленное исследование на основе двух- и трехпараметрических моделей А. Бирнбаума.

Существенным недостатком модели Г. Раша являются ограничения, накладываемые на крутизну характеристических кривых. В частности, она считается одинаковой для всех кривых. Этот недостаток заметен, когда необходимо отдать предпочтение одному из заданий равной трудности. Удаление из теста заданий с более крутыми характеристическими кривыми может привести к снижению надежности и валидности теста.

При анализе тестов, в состав которых включены задания закрытой формы, не рекомендуется использовать двухпараметрическую модель А. Бирнбаума, так как существует заметное отклонение эмпирических данных от теоретической кривой, предсказывающей вероятность правильного выполнения задания при различных значениях переменной $\boldsymbol{\theta}$. Особенно этот эффект характерен для испытуемых с низкими значениями параметра $\boldsymbol{\theta}$ при ответах на трудные задания теста вследствие возможного угадывания правильного ответа. При этом эффект угадывания существенно снижает дифференцирующую способность заданий теста.

В качестве общих недостатков моделей современного метода тестирования при оценке качества тестовых материалов можно рассматривать следующее: для получения достоверных результатов необходимо накопление больших объемов статистической информации; сложный математико-статистический аппарат требует разработки специальных программных продуктов; результаты анализа представляются в виде набора графиков и гистограмм, для их интерпретации необходимы углубленные знания в области тестологии.

Экспертные методы оценки качества тестовых материалов, как правило, используются для уточнения содержания и спецификации тестов и тестовых заданий. Для получения экспертных оценок используют два подхода.

Первый подход предполагает получение экспертных оценок после эмпирического опробования тестовых заданий самими экспертами с последующей обработкой результатов. По результатам экспертизы, например, могут оцениваться: время выполнения теста экспертами, коэффициент решаемости задания экспертами, коэффициент автоматизации (коэффициент навыка) и др.

При использовании второго подхода характеристики тестовых материалов определяются экспертами умозрительно на основе предполагаемого числа и характера умственных операций, необходимых для успешного выполнения задания тестирующимися. При этом обработка мнений экспертов выполняется в соответствии с используемыми экспертными методами. В качестве уточняемых характеристик в этом случае могут выступать: уровень трудности тестового задания (уровень базовости), уровень значимости тестового задания, ступень абстракции и др.

В основе экспертных методов лежат следующие положения. Во-первых, экспертная оценка имеет вероятностный характер и основывается на способности эксперта давать информацию-оценку в условиях неопределенности. Во-вторых, считается, что когда оценку дает не один, а несколько экспертов, то истинное значение исследуемой характеристики находится внутри диапазона оценок отдельных экспертов, т.е. обобщенное коллективное мнение является более достоверным. Обобщенное мнение экспертов можно получить, вычислив элемент результирующего вектора как взвешенное среднее:

$$\lambda_i = \sum_{j=1}^{m} \nu_j \lambda_{ij} \tag{6}$$

где $\lambda_{ij}$ – относительная значимость $\underline{\boldsymbol{i}^{\boldsymbol{zo}}}$ параметра оценки, по мнению $\underline{\boldsymbol{j}^{\boldsymbol{zo}}}$ эксперта, $\nu_j$ – нормированный вес каждого $\underline{\boldsymbol{j}^{\boldsymbol{zo}}}$ эксперта.

Определить вес каждого эксперта можно на основании коэффициента компетентности или на основе анализе отклонений его оценок[6].

Весовой коэффициент $\underline{\boldsymbol{j}^{\boldsymbol{zo}}}$ эксперта обратно пропорционален его вкладу в общую дисперсию оценок:

$$\nu_j = \frac{2}{m} - \frac{s_j}{\sum_{j=1}^{m} s_j} \tag{7}$$

где $s_j = \sum_{i=1}^{n} \sum_{k=1}^{m} (\lambda_{ij} - \lambda_{ik})^2$ – сумма квадратичных отклонений оценок $\underline{\boldsymbol{j}^{\boldsymbol{zo}}}$ эксперта от остальных оценок.

Общее мнение экспертов по всем параметрам оценки ( $\lambda_i^*$ ) можно определить как их среднее значение всех относительных значимостей:

$$\lambda_i^* = \sum_{i=1}^{n} \lambda_i \tag{8}$$

В-третьих, обработка полученных оценок проводятся по определенному алгоритму. Затем отобранные и подготовленные эксперты действуют в соответствии с разработанными правилами.

Важнейшим показателем качества экспертизы выступает согласованность оценок экспертов, которая показывает, насколько близки или далеки друг от друга точки зрения экс-

пертов. Синтез обобщенного мнения осуществляется ***статистическим*** или ***алгебраическим*** методами.

При использовании ***статистического*** метода определяют отклонение оценок экспертов, используя значения среднего арифметического, среднего взвешенного, суммы рангов, мажоритарной выборки. При использовании ***алгебраического*** метода оценивают качество результирующей экспертной оценки (т.е. той оценки, которая наименее всего отстоит от остальных) с помощью медианы распределения, медианы Кемени, принципа Кондорсе. Вследствие сложного математического аппарата обработки полученных результатов экспертизы необходимо дополнительное специальное программное обеспечение.

При оценке качества тестовых материалов (разработанных на базе учебных заведений) с использованием экспертных методов существенной оказывается проблема, связанная с определением численности экспертной группы. Как правило, в качестве экспертов выступают преподаватели того же вуза, являющиеся либо специалистами соответствующей предметной области, либо родственной с ней. Однако тестовые материалы разрабатываются по конкретным (узко направленным) разделам учебной дисциплины. Вследствие существующей специфики отдельных дисциплин и дефицита в вузах специалистов по этим дисциплинам возникает проблема с подбором экспертов. В результате состав экспертной группы оказывается малочисленным и не позволяет обеспечивать достаточную статистическую достоверность выборочной оценки, так как при небольшом числе экспертов на общую групповую оценку существенное влияние оказывают индивидуальные оценки экспертов.

С целью увеличения численности экспертной группы было бы рациональным привлечение внешних специалистов (например, преподавателей других вузов), однако при этом возникают дополнительные организационные трудности проведения экспертного исследования, увеличиваются затраты времени и денежных средств на проведение экспертизы.

Рассмотренные методы оценки качества тестовых материалов обладают рядом существенных недостатков, которые сужают область их применения. На сегодняшний день не существует оптимального метода, позволяющего провести точную оценку качества тестовых материалов.

---

[1] ***Глас Дж.***, ***Стенли Дж.*** Статистические методы в педагогике и психологии. М.: Прогресс, 1976. 495 с.

[2] ***Поддубная Л.М.*** Компьютерная технология разработки тестовых заданий: Учеб. пособие. М.: Логос, 2003. 56 с.

[3] ***Keeves J.P.*** (Ed.) Educational Research, Methodology and Measurement: An International Handbook. Oxford, Pergamon Press, 1988.

[4] ***Нейман Ю.М.***, ***Хлебников В.А.*** Введение в теорию моделирования и параметризации педагогических тестов. М.: Прометей, 2000. 168 с.

[5] ***Челышкова М.Б.*** Теория и практика конструирования педагогических тестов: Учеб. пособие. М.: Логос, 2002. 432 с.

[6] ***Анохин А.Л.*** Методы экспертных оценок: Учеб. пособие. Обнинск: ИАТЭ, 1996. 148 с.