УДК 651.3:518.5

П.І. ФЕДОРУК

АДАПТИВНІ ТЕСТИ: СТАТИСТИЧНІ МЕТОДИ АНАЛІЗУ РЕЗУЛЬТАТІВ ТЕСТОВОГО КОНТРОЛЮ ЗНАНЬ

Abstract: *In the article the possibilities of static methods of analysis of test results, given are considered and are the simplest and the most necessary procedures of statistical processing of knowledge test results and methods of evaluation of test quality. The article describes the approach to test construction which is described in the modern test theory on the basis of mathematical theory of parametric evaluation of test tasks on the basis of modern mathematical models: oneparametric model of Rush, twoparametric and threeparametric models of Birnbaum.*
Key words: *adaptive test, statistical processing, parametric evaluation.*

Анотація: *У статті розглянуто можливості статистичних методів аналізу результатів тестування, представлено найпростіші та необхідні процедури статистичної обробки результатів тестування знань і методи оцінки якості тесту. Розглянуто підхід до конструювання тестів, представлений у сучасній теорії тестування на основі математичної теорії параметричної оцінки тестових завдань на базі основних сучасних математичних моделей: однопараметричної моделі Раша, двопараметричної та трипараметричної моделях Бірнбаума.*
Ключові слова: *адаптивний тест, статистична обробка, параметрична оцінка.*

Аннотация: *В статье рассмотрены возможности статистических методов анализа результатов тестирования, приведены простейшие и необходимые процедуры обработки результатов тестирования знаний и методы оценки качества теста. Рассмотрен поход к конструированию тестов, который представлен в современной теории тестов на основе математической теории параметрической оценки тестовых заданий на базе основных современных моделей: однопараметрической модели Раша, двухпараметрической и трехпараметрической моделях Бирнбаума.*
Ключевые слова: *адаптивный тест, статистическая обработка, параметрическая оценка.*

1. Вступ

З розвитком освітніх систем все більше уваги приділяється контролю знань тих, хто навчається за допомогою тестування. Сучасне тестування являє собою комплекс стандартизованих методів вимірювання тих латентних (тобто недоступних для безпосереднього спостереження) параметрів людини, які визначають її рівень підготовки і відповідність освітнім стандартам у конкретній області знань [1, 2]. При цьому широко використовуються математичні методи планування й обробки результатів тестування, а також сучасні технології обробки інформації. Об'єктивний контроль знань, вмінь і навичок – одне із актуальних завдань нашого часу [3]. Його вдається виконати при критеріально-орієнтованій інтерпретації тестування. Критеріально-орієнтоване тестування призначене не тільки для оцінювання рівня знань, а й для визначення рівня індивідуальних досягнень відносно певного критерію на підставі логіко-функціонального аналізу змісту завдань. Тому, враховуючи індивідуалізацію навчання, конструювання критеріально-орієнтованих тестів є одним із провідних та найактуальніших напрямків розвитку теорії тестів.

В сучасних навчальних системах тест повинен бути індивідуалізований. Тобто, він повинен мати певну довжину, а для всіх його завдань, апробованих емпірично, необхідно однозначно знати їх складність [3]. Так виникає одне із головних питань теорії тестів – питання побудови оптимального тесту.

2. Класичні статистичні методи аналізу результатів тестування

Історично виділяють два основні підходи до створення тестів. Перший з них набув широкого розвитку в рамках класичної теорії тестів. Згідно з ними, рівень знань учасників тестування

оцінюється за допомогою їх індивідуальних балів. Бал обчислюють як алгебраїчну суму оцінок виконання кожного завдання тесту.

Класична теорія тестів ґрунтується на статистичних методах аналізу результатів тестування [4, 5]. Розглянемо найпростіші і необхідні процедури статистичної обробки результатів тестування знань і методи оцінки якості тесту.

В усіх відомих теоріях тестування розглядається як процес протистояння учасника із запропонованими йому завданнями. Позначимо через x_{ij} числову оцінку успішності виконання j-ого завдання i-им студентом. Результати тестування звичайно представляються у вигляді матриці $\{x_{ij}\}$ з n рядками та m стовпцями ($i=\overline{1,n}$, $j=\overline{1,m}$). Матриця тестових результатів показує результат виконання всіх завдань учасниками тестування. На практиці прийнято, як правило, використовувати дихотомічну шкалу оцінок результатів. У результаті правильного виконання завдання студент отримує один бал, $x_{ij}=1$, в протилежному випадку – нуль балів, $x_{ij}=0$.

Якщо за правильне виконання завдання студент отримує одиницю, а за неправильне – нуль, то бал виражає кількість правильно виконаних завдань. Результат можна оцінювати не лише нулем чи одиницею, але й присвоювати певний ваговий коефіцієнт, що відповідає складності завдання.

Процес статистичної обробки матриці результатів тестування будемо розглядати послідовно. На першому кроці обчислюємо індивідуальні початкові бали всіх студентів $y_i, i=\overline{1,n}$.

$y_i=\sum_{j=1}^{m} x_{ij}$ – результат (індивідуальний бал) i-ого студента після проходження тесту (кількість усіх правильних відповідей).

Обчислюємо середній результат $\overline{y}$ сумарних балів учасників тестування та середній результат $\overline{x}_j$ студентів за кожним завданням:

$$\overline{y}=\frac{\sum_{i=1}^{n} y_i}{n}, \quad \overline{x}_j=\frac{\sum_{i=1}^{n} x_{ij}}{n}.$$

Важливою вимогою до тестових завдань є їх об'єктивний рівень складності. В тесті немає місця завданням з невідомою мірою складності. Завдання стають тестовими лише після емпіричної перевірки міри їх складності.

Складність завдань можна визначати двома способами [6]:

– на основі оцінки передбачуваного числа і характеру розумових операцій, необхідних для вдалого виконання завдань;

– на основі емпіричної перевірки завдань, з підрахунком частки неправильних відповідей.

У класичній теорії тестів багато років розглядалися тільки емпіричні показники складності. У сучасних теоріях навчальних тестів, які використовуються в дистанційному навчанні, більше уваги приділяється характеру розумової діяльності у процесі виконання тестових завдань різних форм.

Емпірично складність завдання визначається додаванням елементів матриці по рядках і дорівнює числу правильних відповідей, отриманих за кожним завданням (R_j). Чим більше правильних відповідей на завдання, тим воно легше для даної групи студентів.

У міру простоти показник R_j зручний, але доти, поки не з'являться інші групи з іншим числом студентів. Тому для одержання об'єктивних характеристик R_j ділять на число студентів у кожній групі (об'єм вибірки):

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

У результаті отримаємо нормований статистичний показник – частка правильних відповідей, p_j. Статистика p_j довго використовувалася як показник рівня складності завдання в класичній теорії тестів. Пізніше була усвідомлена певна її неточність: адже збільшення значення p_j означає не зростання складності завдання, а, навпаки, зростання легкості. Тому з показником складності завдань стали асоціювати протилежну статистику – частку неправильних відповідей, q_j. Вона обчислюється як відношення числа неправильних відповідей W_j (від англ. wrong – неправильний) до кількості учасників тестування n:

$$q_j = \frac{W_j}{n},\ p_j + q_j = 1.$$

Наступною вимогою до тестових завдань є варіація балів.

Якщо на деяке завдання правильно відповідають всі студенти, то таке завдання стає не тестовим. Учасники тестування відповідають на нього однаково: між ними немає варіації. Відповідно по даному завданню в матриці будуть стояти лише одиниці. Нетестовим вважається завдання, на яке немає жодної правильної відповіді. Варіація по ньому теж рівна нулю. Нульова варіація означає практичну необхідність викидання завдання із тесту.

Зручною мірою варіації є значення дисперсії s_y^2 і стандартне відхилення s_y сумарних балів учасників тестування:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1},\ s_y = \sqrt{s_y^2}$$

та величина s_j^2 – дисперсія результатів студентів по j-ому завданню:

$$s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1},\ j = \overline{1,m}.$$

Якщо успішність виконання j-ого завдання оцінюється балами 0 чи 1, то міра варіації визначається формулою $s_j^2 = p_j(1-p_j)$ або $s_j^2 = p_j \cdot q_j$.

Обчисливши дисперсію, можна знайти і стандартне відхилення $s_j = \sqrt{s_j^2}$.

Завдання в тестовій формі не можна назвати тестовим, якщо воно не корелює із сумою балів по всьому тесту. Для цього можна використовувати коефіцієнт кореляції Пірсона:

$$R_j = \frac{\dfrac{\sum_{i=1}^{n}(x_{ij} \times y_i)}{n} - \overline{x}_j \times \overline{y}}{s_j \times s_y} \times \frac{n}{n-1}$$

або бісеріальний коефіцієнт кореляції:

$$B_j = \frac{M_{j1} - M_{j0}}{s_y} \cdot \sqrt{\frac{n_{j0} \cdot n_{j1}}{n(n-1)}},$$

де $n_{j1} = \sum_{i=1}^{n} x_{ij}$ – число тих студентів, що одержали за даним завданням 1 бал;

$n_{j0} = n - n_{j1}$ – число тих студентів, що відповіли неправильно на j-е завдання;

M_{j1} – середнє арифметичне сум балів по всьому тесту для тих студентів, які одержали за даним завданням 1 бал, M_{j0} – нуль балів:

$$M_{j0} = \frac{\sum_{i=1}^{n}(1 - x_{ij}) y_i}{n_{j0}}, \quad M_{j1} = \frac{\sum_{i=1}^{n} x_{ij} y_i}{n_{j1}}.$$

Попарний кореляційний зв'язок завдань між собою можна обчислити за формулою

$$\varphi_{jk} = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}},$$

де $A = \sum_{i=1}^{n} x_{ij} x_{ik}$ – кількість учасників тестування, які вірно виконали завдання j та k;

$B = \sum_{i=1}^{n} x_{ij}(1 - x_{ik})$ – вірно виконали завдання j та невірно k.

Аналогічно $C = \sum_{i=1}^{n}(1 - x_{ij}) x_{ik}$, $D = \sum_{i=1}^{n}(1 - x_{ij})(1 - x_{ik})$.

Із збірника завдань викидаємо завдання, що не володіють дискримінативністю: $p_j > 0{,}9$ (надто легкі), $p_j < 0{,}2$ (надто важкі). Виключають завдання, що погано корелюють із сумою балів ($B_j < 0{,}15$), і негативні коефіцієнти кореляції. Для зменшеного списку завдань складається нова впорядкована таблиця, для якої перераховуються вищезгадані показники.

Крім того, отримані тестові завдання повинні задовольняти критерій надійності та валідності.

Надійність тесту тим вища, чим більше погоджені результати учасника тестування при повторній перевірці знань за допомогою того самого тесту. Погодженість можна вимірювати коефіцієнтом надійності Кьюдера-Річардсона:

$$\rho = \frac{s_y^2 - \sum_{j=1}^{m} s_j^2}{2s_y^2} + \sqrt{\left(\frac{s_y^2 - \sum_{j=1}^{m} s_j^2}{2s_y^2} \right)^2 + \frac{\sum_{j=1}^{m} B_j^2 s_j^2}{2s_y^2}}.$$

Якщо $0{,}8 \le \rho \le 0{,}89$, тест має високу надійність, якщо $\rho \ge 0{,}9$, надійність дуже висока. Чим вищий показник надійності, тим менша помилка виміру індивідуального результату.

Валідність тесту показує, наскільки якісно робить тест те, для чого він був створений. Визначити коефіцієнт валідності тесту – означає визначити, як виконання тесту співвідноситься з іншими незалежно зробленими оцінками знань учасників тестування. Для визначення валідності необхідним є незалежний зовнішній критерій, тобто оцінка експерта (викладача). За коефіцієнт валідності приймають коефіцієнт кореляції результатів тестових вимірів і критерію. Якщо експертна оцінка знань студентів, отримана незалежно від процедури тестування, представлена числовою послідовністю $Y_1, Y_2, ..., Y_n$, то коефіцієнт валідності тесту може бути обчислений так:

$$V = \frac{\frac{\sum_{i=1}^{n} (Y_i \cdot y_i)}{n} - \overline{Y} \cdot \overline{y}}{s_Y \cdot s_y} \cdot \frac{n}{n-1},$$

де $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, $s_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}}$ – стандартне відхилення експертних оцінок.

3. Математична теорія параметричної оцінки тестових завдань

Штучність низки припущень класичної теорії тестів і деякі її практичні недоліки помітно вплинули на ріст критичних тенденцій. Цьому, в першу чергу, сприяли сумніви в об'єктивності емпіричних оцінок складності завдань тесту. А саме: виникло питання про правомірність традиційного оцінювання складності завдань за допомогою частки правильних чи неправильних відповідей.

При традиційному підході до зміни рівня складності завдань на різних по підготовці вибірках студентів залишається відкритим питання про об'єктивність значень параметра складності завдань тесту [1]. Спроба введення вагових коефіцієнтів, що відображають вклад завдання в індивідуальний бал студента, суттєво не виправляє такі недоліки. Значення цих коефіцієнтів можна, в свою чергу, поставити під сумнів. Деякі з них визначаються суб'єктивно, на основі думки педагога про складність завдання. Оцінки решти з них базуються на емпіричних даних тестування і, відповідно, залежать від рівня знань вибірки студентів.

Таким чином, можна відзначити, що нестійкість статистик та їх взаємний вплив помітно знижують якість тестових результатів. За допомогою цих статистик не можна об'єктивно оцінити значення параметрів, що характеризують складність завдання тесту, а також виразити значення цих параметрів на інтервальній шкалі [2].

Другий підхід до створення тестів та обробки їх результатів представлений в так званій сучасній теорії тестування, що набула широкого розвитку в 1960 – 1980 роках в багатьох західних країнах [7].

Сучасний етап розвитку і функціонування тестового контролю характеризується застосуванням до вирішення психолого-педагогічних задач методології латентно-структурного аналізу (LSA) [8]. Одним з напрямків LSA є Item Response Theory (IRT) – математична теорія параметричної оцінки тестових завдань і тих, хто проходить тестування. Відповідно до цієї теорії встановлено, що між результатом виконання, що спостерігається, і латентним параметром учасників тестування є деяка залежність, яку можна виразити за допомогою функції. Для IRT характерне прагнення до фундаментального теоретичного підходу і разом з цим до коректного розв'язання низки практичних задач.

IRT спрямована на оцінювання латентних якостей особистості та параметрів завдань тесту на основі математичних моделей [8, 9].

До найбільш вагомих переваг IRT відносять:

– стійкі об'єктивні оцінки параметра складності завдань, що не залежать від властивостей вибірки студентів, які виконують тест;

– вимірювання значень параметрів студентів і завдань тесту в одній і тій же шкалі, що дозволяє поставити у відповідність рівень знань кожного учасника тестування із рівнем складності кожного завдання тесту;

– можливість оцінити ефективність різних за рівнем складності завдань для вимірювання даного значення латентного параметра студента.

На відміну від класичної теорії тестів, де індивідуальний бал розглядається як стале число, в IRT латентний параметр трактується як деяка змінна. Початкове значення параметра отримується безпосередньо на основі емпіричних даних тестування. Змінний характер вимірюваної величини вказує на можливість послідовного наближення до об'єктивних оцінок параметрів за допомогою ітераційних методів.

Латентні параметри, точніше, взаємодія двох множин їх значень породжує результати виконання тесту. Елементи першої множини – це значення латентного параметра, що визначає рівень знань n учасників тестування θ_i, де $i = 1,...,n$. Другу множину утворюють значення латентного параметра δ_j, де $j = 1,...,m$, що відповідають рівням складності m завдань тесту.

На практиці ставиться задача: за відповідями студентів на завдання тесту оцінити значення латентних параметрів θ і δ [10]. Для її вирішення потрібно відповісти на два питання:

1. Як вибрати співвідношення між θ і δ ?

2. Як правильно вибрати математичну модель, тобто таку модель, яка пов'язує емпіричні результати тестування та латентні параметри θ і δ?

В рамках IRT датським математиком Джорджем Рашем у 1957 році була запропонована модель контролю знань [11], яку часто називають простою логістичною моделлю. Модель Раша спирається на поняття „складність завдання" та „рівень підготовки студентів". Так, одне завдання вважається складнішим, ніж друге, якщо ймовірність правильної відповіді на перше завдання менша, ніж на друге, незалежно від того, хто його виконує.

Таким чином, оцінка складності тестових завдань не залежить від вибірки учасників тестування. Крім того, модель Раша характеризується найменшим числом параметрів: один параметр рівня знань для всіх випробуваних та тільки один параметр складності для всіх завдань.

Дж. Раш запропонував ввести співвідношення між θ і δ у вигляді різниці $\theta-\delta$, вважаючи, що параметри θ і δ оцінюються в одній шкалі.

У такій математичній моделі параметри θ і δ виражаються як показники, задані в одній шкалі логітів. Введення однієї шкали для елементів двох множин θ і δ дозволяє ввести взаємозв'язок між змінними у вигляді різниці $\theta-\delta$, коректно порівняти результати студентів, отримані за допомогою різних тестів, оцінити рівень складності завдань незалежно від рівня підготовки груп студентів.

Можна розглядати умовну ймовірність правильного виконання j-ого завдання із рівнем складності δ_j різними студентами. Тут незалежною змінною є θ, а δ_j – параметр, що визначає складність j-ого завдання:

$$P_j\left(x_{ij}=1\middle|\,\delta_j\right)=\varphi\left(\theta-\delta_j\right),\ j=1,...,m.$$

В теорії IRT функцію $\varphi(\theta)$ називають „Item response function" (IRF). Спеціальну назву має графік такої функції – характеристична крива j-ого завдання (ICC). При виборі вигляду функції P_j враховують обставини як емпіричного, так і математичного характеру. Припустивши, що значення латентних параметрів змінних θ і δ мають нормальний розподіл, маємо дві такі функції. Одна з них позначається $\Psi(x)$ – деяка логістична функція, інша $\Phi(x)$ – інтегральна функція нормованого нормального розподілу. Оскільки для одних і тих же значень x ординати точок графіків функцій $\Phi(x)$ і $\Psi(1,7x)$ відрізняються достатньо мало, а саме

$$\left|\Phi(x)-\Psi(1,7x)\right|<0,01,$$

то на практиці перевагу віддають функції $\Psi(1,7x)$, адже в ній значно простіше аналітичне завдання, вигідне для оцінювання δ.

Кількість параметрів у такому аналітичному завданні функції розбиває сімейства IRF на класи. Серед логістичних функцій розрізняють:

– однопараметричну модель Дж. Раша:

$$P_j(\theta)=\frac{e^{1,7(\theta-\delta_j)}}{1+e^{1,7(\theta-\delta_j)}};$$

– двопараметричну модель А. Бірнбаума:

$$P_j(\theta) = \frac{e^{1,7a_j(\theta-\delta_j)}}{1+e^{1,7a_j(\theta-\delta_j)}};$$

– трипараметричну модель А. Бірнбаума:

$$P_j(\theta) = c_j + (1-c_j)\frac{e^{1,7a_j(\theta-\delta_j)}}{1+e^{1,7a_j(\theta-\delta_j)}},$$

де a_j та c_j – другий і третій параметри, що відповідно характеризують диференційовану здатність завдання при зміні різних значень θ та ймовірність угадування правильної відповіді на j-е завдання.

У кожній із представлених моделей параметри θ і δ виражаються як показники єдиної для всіх моделей шкали логітів. Введення єдиної шкали логітів для елементів цих двох різних множин дозволяє ввести взаємозв'язок між змінними у вигляді різниці, оцінити складність завдань тесту незалежно від рівня підготовки груп учасників тестування.

4. Основні математичні моделі

Модель Раша

Успіх учасника тестування при розв'язанні деякого тестового завдання залежить від двох факторів: складності завдання і рівня підготовки учасника. Ймовірність того, що деякий учасник вірно виконає конкретне завдання, є функцією щонайменше двох аргументів: рівня підготовки учасника тестування S та рівня складності даного завдання t:

$$P = P(S,t).$$

Таку функцію називають функцією успіху [8]. Якщо вигляд функції успіху відомий, то за результатами випробувань методами математичної статистики з певною точністю можна оцінити аргументи цієї функції, в тому числі і рівень складності завдань [8].

Основна логістична модель Раша полягає в тому, що ймовірність правильної відповіді першим учасником (більш підготовленим) на перше завдання має співпадати із ймовірністю правильного виконання другим учасником (менш підготовленим) іншого завдання (менш складного).

Із цього слідує:

– аргументи S і t тісно пов'язані між собою, неможливо визначити один із них, не визначивши другий;

– ймовірність успіху залежить не від кожного аргументу S і t окремо, а від їх відношення

$$P = P(S,t) = P_1(\xi), \quad \xi = \frac{S}{t}.$$

Параметри S і t називають латентними (неспостережуваними) параметрами [10], оскільки вони описують деякі приховані характеристики учасників тестування та тестових завдань.

Функція успіху запишеться у вигляді однорідної функції нульового порядку :

$$P = P(S,t) = \frac{S}{S+t} = \frac{S/t}{S/t+1} = \frac{\xi}{\xi+1}.$$

Ця найпростіша модель вперше дала можливість об'єктивно визначати співвідношення між учасниками тестування і тестовими завданнями довільних рівнів підготовки та складності.

Параметри S і t можуть бути довільними додатніми числами $S \in (0,\infty)$, $t \in (0,\infty)$.

Якщо ввести позначення

$$\ln S = 1{,}7 \cdot \theta,\ \ln t = 1{,}7 \cdot \delta \ \Leftrightarrow\ S = e^{1{,}7\theta},\ t = e^{1{,}7\delta},\ \ln \xi = \ln S - \ln t = 1{,}7(\theta - \delta),$$

то функція успіху матиме вигляд

$$P = \frac{e^{1{,}7\theta}}{e^{1{,}7\theta} + e^{1{,}7\delta}} = \left(1 + e^{1{,}7(\delta-\theta)}\right)^{-1} = \left(1 + e^{-1{,}7(\theta-\delta)}\right)^{-1}$$

і буде називатися основною логістичною моделлю Раша.

Ймовірність успіху залежить лише від різниці $\theta - \delta$, і тому модель Раша є однопараметричною.

Розглянемо модель Раша більш детально. Нехай тест складається із m різних завдань, тест виконують n студентів. Позначимо через x_{ij} числову оцінку успішності виконання j-ого завдання i-им студентом. Якщо i-ий студент вірно виконав j-те завдання, то $x_{ij} = 1$. Якщо невірно, то $x_{ij} = 0$. Результати тестування представляються у вигляді матриці результатів $\{x_{ij}\}$, де $i = \overline{1,n}$, $j = \overline{1,m}$.

Обчисливши $p_i = \dfrac{\sum_{j=1}^{m} x_{ij}}{m}$ (частка правильних відповідей i-ого студента на всі завдання тесту та $q_i = 1 - p_i$ – частка неправильних відповідей), можна визначити початковий логіт рівня знань кожного студента (тобто початкову оцінку рівня знань i-ого студента у шкалі логітів):

$$\theta_i^0 = \ln \frac{p_i}{q_i},\ i = \overline{1,n}\ .$$

Обчисливши $p_j = \dfrac{\sum_{i=1}^{n} x_{ij}}{n}$ (частка правильних відповідей всіх студентів групи на j-е завдання та $q_j = 1 - p_j$ – частка неправильних відповідей), можна визначити початковий логіт складності завдання (тобто початкову оцінку рівня складності j-ого завдання у шкалі логітів):

$$\delta_j^0 = \ln \frac{q_j}{p_j},\ j = \overline{1,m}.$$

Цей етап оцінювання латентних параметрів є початковим. Після його завершення кожен із параметрів буде виражений в інтервальній шкалі, але з різними значеннями середнього та різними стандартними відхиленнями.

На наступному етапі значення θ_i^0 та δ_j^0 переводимо в одну інтервальну шкалу. У формулі для такого переходу закладена ідея зниження впливу складності завдань на оцінки учасників тестування.

Попередньо обчисливши середнє значення початкових логітів рівня знань студентів:

$$\overline{\theta} = \frac{\theta_1^0 + ... + \theta_n^0}{n}$$

та стандартне відхилення V розподілу початкових значень параметра δ:

$$V^2 = \frac{\sum_{i=1}^{n} (\theta_i^0 - \overline{\theta})^2}{n-1}, \; V = \sqrt{V^2},$$

отримаємо формулу для обчислення логіта складності j-ого завдання:

$$\delta_j = \overline{\theta} + Y \cdot \delta_j^0, \; j = \overline{1,m},$$

де $Y = \left(1 + \frac{V^2}{2,89}\right)^{1/2}$.

Аналогічно, обчисливши $\overline{\delta} = \frac{\delta_1^0 + ... + \delta_m^0}{m}$, $W = \sqrt{\frac{\sum_{j=1}^{m} (\delta_j^0 - \overline{\delta})^2}{m-1}}$ та $X = \left(1 + \frac{W^2}{2,89}\right)^{1/2}$,

отримаємо формулу для обчислення логіта рівня знань i-ого студента:

$$\theta_i = \overline{\delta} + X \cdot \theta_i^0, \; i = \overline{1,n}.$$

Така оцінка параметра δ_j дозволяє оцінити рівень складності всіх завдань незалежно від рівня підготовки студентів. Теоретично значення параметра δ змінюються на інтервалі $(-\infty;\infty)$, але на практиці рекомендованим є інтервал $(-3;3)$.

Отримані значення дозволяють співставити рівень знань студентів із рівнем складності завдань тесту. Якщо $\theta_i - \delta_j$ – від'ємна величина і велика за модулем, то завдання складності δ_j є надто важким для студента з рівнем знань θ_i, і воно не буде корисним для виміру рівня знань i-ого студента. Якщо ця різниця додатня і велика за модулем, то завдання надто легке, воно давно освоєно студентом. Якщо $\theta_i = \delta_j$, то ймовірність того, що студент вірно виконає завдання, дорівнює 0,5.

За допомогою різниці $\theta_i - \delta_j$ визначається планка виконання тесту, тобто ймовірність, яка стверджує: може студент пройти тест чи ні.

Після оцінювання значень θ і δ у шкалі логітів приступають до обчислення ймовірності $P_j(\theta)$ правильного виконання j-ого завдання тесту різними студентами:

$$P_j(\theta) = \frac{e^{1,7(\theta-\delta_j)}}{1+e^{1,7(\theta-\delta_j)}},$$

де $\theta = (\theta_1, \theta_2, ..., \theta_n)$.

Ймовірність P_j правильного виконання j-ого завдання тесту є зростаючою функцією змінної θ. Очевидно, що чим вищий рівень знань студента, тим більша ймовірність правильного виконання ним j-ого завдання тесту.

Ввівши умовну ймовірність P_j правильного виконання j-ого завдання різними студентами, можна перейти до побудови характеристичної кривої j-ого завдання тесту (рис. 1).

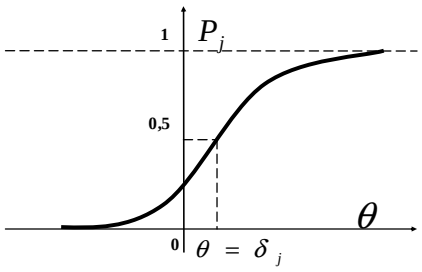

Рис. 1. Характеристична крива j-ого завдання

Характеристична крива j-ого завдання тесту показує взаємозв'язок між значеннями незалежної змінної θ і значеннями P_j [8]. Точці перегину характеристичної кривої відповідає значення $\theta = \delta_j$, а P_j в цій точці дорівнює $0,5$. Таким чином, студент із рівнем знань, що дорівнює складності j-ого завдання тесту, відповість на нього правильно з ймовірністю $0,5$. Для студентів з рівнем знань набагато більшим, ніж δ_j, ймовірність правильної відповіді на це завдання прямує до одиниці. Якщо ж значення θ розміщені достатньо далеко від значення $\theta = \delta_j$ і зліва від точки перегину, то ймовірність правильного виконання j-ого завдання буде прямувати до нуля.

Характеристичні криві, що відповідають завданням різних рівнів складності, не перетинаються [8], [10].

Збільшення складності j-ого завдання тесту на константу C $(C > 0)$ зумовить зміщення характеристичної кривої вправо. Із попередньою ймовірністю на це завдання буде відповідати студент із рівнем знань $\theta + C$. Оскільки $\theta - \delta = (\theta + C) - (\delta + C)$, то значення функції $P_j(\theta)$ не змінюється.

Отже, якщо взяти важче завдання, то з колишньою ймовірністю на нього буде відповідати той студент, у якого рівень підготовки зміниться на ту ж константу, що і рівень складності завдання.

Враховуючи принцип індивідуалізації навчання, для більш грунтовного вирішення необхідний додатковий аналіз учасників тестування [9]. Якщо група студентів гомогенна за рівнем знань і більшість значень θ розміщені на невеликому інтервалі осі латентної змінної, то і більша частина завдань тесту за складністю має відповідати цьому інтервалу. У випадку гетерогенної за рівнем знань вибірки студентів значення параметра складності повинні охоплювати більший інтервал на осі θ, а характеристичні криві завдань можуть розміщуватись достатньо далеко одна від одної.

Одне і те ж завдання може бути як ефективним, так і неефективним при оцінюванні різних значень θ. Тому не існує єдиної оптимальної моделі при підборі завдань у тест. Запропоноване моделювання шляхом цілеспрямованого підбору завдань для оцінювання даного θ_i дозволяє лише мінімізувати стандартну похибку вимірювання його значення.

Двопараметрична модель Бірнбаума

Формулу для умовної ймовірності правильного виконання j-ого завдання тесту учасниками із різними значеннями θ у випадку двопараметричної моделі Бірнбаума можна записати у вигляді

$$P_j\left(x_{ij}=1\middle|\,\delta_j\right)=\frac{e^{1{,}7a_j(\theta-\delta_j)}}{1+e^{1{,}7a_j(\theta-\delta_j)}},$$

де, крім попередніх позначень, вводиться нове a_j для другого параметра j-ого завдання тесту.

Ймовірність успіху залежить лише від a_j та $\theta-\delta_j$, тому модель Бірнбаума називають двопараметричною [8] (див. рис. 2).

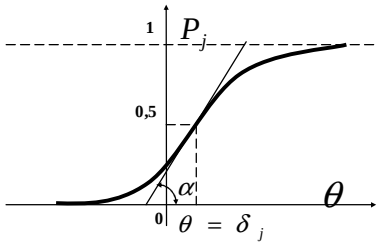

Рис. 2. Характеристична крива j-ого завдання за моделлю Бірнбаума

При геометричному трактуванні перший параметр δ_j можна розглядати як характеристику положення кривої j-ого завдання відносно осі θ. Другий параметр a_j пов'язаний із кривизною характеристичної кривої j-ого завдання в точці її перегину. А саме, значення a_j прямо пропорційне тангенсу кута нахилу дотичної до характеристичної кривої завдання в точці $\theta = \delta_j$. Це означає, що крутіші криві відповідають більшим значенням a_j, відповідно для похилих кривих $a_j \to 0$.

Теоретично значення параметра a_j можуть змінюватись на інтервалі $(-\infty;\infty)$. Аналіз характеристичних кривих завдань однакової складності, але різної кривизни дозволяє відібрати кращі завдання і визначити розумні межі інтервалу для значень параметра a_j [11].

При невеликих значеннях параметра a_j характеристична крива є похилою, тому для учасників тестування із рівнем підготовки $\theta < \delta_j$ і для учасників із $\theta > \delta_j$ ймовірності правильного виконання j-ого завдання приблизно однакові. Якщо значення параметра a_j досить велике, то ймовірності успіху будуть суттєво відрізнятися. Тому параметр a_j отримав назву коефіцієнта дискримінації, тобто числової характеристики здатності тестового завдання диференціювати учасників тестування за їх рівнем підготовки.

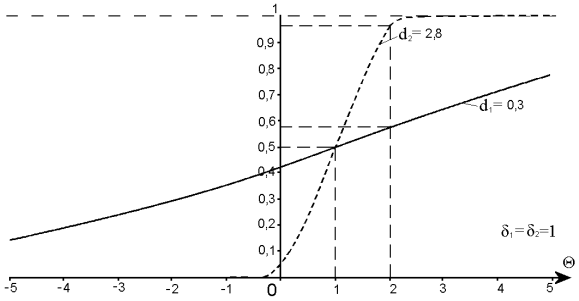

Рис. 3. Характеристичні криві завдань з однаковим рівнем складності $\delta = 1$ і різними коефіцієнтами дискримінації a_j

На рис. 3 зображено характеристичні криві для завдань з однаковим рівнем складності $\delta = 1$, але з різними коефіцієнтами дискримінації $a_1 = 0,3$ та $a_2 = 2,8$. Характеристична крива, яка зображена суцільною лінією, відповідає завданню з коефіцієнтом дискримінації $a_1 = 0,3$, а пунктиром – із $a_2 = 2,8$. Розглянемо точку, що відповідає значенню $\delta = 1$ на осі θ. З графіка видно, що для кривої з $a_1 = 0,3$ ймовірність правильного виконання завдання учасниками з рівнем підготовки $\theta < 1$ та $\theta > 1$ відрізняється несуттєво. Для $\theta = 0$ ймовірність успіху дорівнює 0,43, а для $\theta = 2$ відповідна ймовірність рівна 0,57. Для характеристичної кривої завдання з $a_2 = 2,8$ ймовірність правильного виконання завдання учасниками тестування з рівнем підготовки $\theta < 1$ та $\theta > 1$ відрізняється досить суттєво. Для $\theta = 0$ така ймовірність рівна 0,06, а для $\theta = 2$ відповідно 0,94.

Відбір завдань із великими значеннями a_j є одним із важливих принципів при підготовці ефективного тесту. Адже при індивідуалізованому підході до навчання немає потреби давати сильному студенту завдання з від'ємними значеннями a_j. На такі завдання відповідають правильно з великою ймовірністю учасники тестування з низьким рівнем знань, а для студентів із високим рівнем знань ймовірність правильної відповіді прямує до нуля. Порівняльний аналіз кривизни характеристичних кривих із спільною точкою перегину дозволяє виділити одне, найбільш ефективне завдання з найбільшим значенням коефіцієнта дискримінації a_j. На практиці, як

правило, рекомендується залишати завдання зі значеннями $a_j \in (0,5;3)$. Значення $a_j = 1$ відповідає однопараметричній моделі Раша.

Трипараметрична модель Бірнбаума

Для тестів з завданнями в закритій формі інколи спостерігається суттєве відхилення емпіричних даних від теоретичної кривої, що характеризує ймовірність правильного виконання завдання при різних значеннях параметра θ. Такий ефект найбільш характерний для учасників з низькими значеннями параметра θ при відповідях на найскладніші завдання тесту. Спроби з'ясувати причини такого відхилення привели творців сучасної теорії тестів до висновку про вплив ефекту вгадування правильної відповіді на достовірність емпіричних даних [8, 9].

Можливо, що учасники тестування із різним рівнем знань користуються різними методами при виборі правильної відповіді. Точніше, методами користуються тільки ті з них, хто володіє достатніми знаннями для правильного вибору. Інші ж, знання яких характеризуються низькими значеннями параметра θ, просто вгадують правильну відповідь на завдання. І чим складніше завдання, тим ймовірніше, що відповідь одержана саме таким чином. Для того, щоб врахувати фактор вгадування, А. Бірнбаум запропонував трипараметричну логістичну модель [11].

У такому випадку ймовірність правильної відповіді студентом на j-е завдання тесту знаходять за формулою

$$P_j(\theta) = P_j(x_{ij} = 1 \big| \delta_j) = c_j + (1 - c_j)\frac{e^{1,7a_j(\theta - \delta_j)}}{1 + e^{1,7a_j(\theta - \delta_j)}},$$

де, крім попередніх позначень, введено третій параметр c_j, що характеризує ймовірність правильної відповіді учасником тестування на j-е завдання тесту при відсутності знань у студента, тобто c_j – це ймовірність вгадування правильної відповіді на j-е завдання. Наприклад, для завдання з п'ятьма варіантами відповідей за класичним означенням $c_j = 0,2$, а з чотирма запропонованими відповідями $c_j = 0,25$.

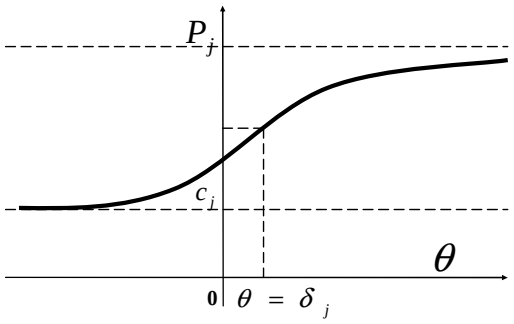

Рис. 4. Характеристична крива j-ого завдання тесту для трипараметричної моделі

Графік характеристичної кривої j-ого завдання у випадку трипараметричної моделі зображено на рис. 4.

Цікаво порівняти кривизну кривої на рисунку із зображеною характеристичною кривою завдання, що має ту ж точку перегину, але нижньою асимптотою якої є вісь θ $(c_j = 0)$. На основі такого порівняння легко бачити, що наявність

третього параметра $(c_j \neq 0)$ перетворює характеристичну криву на більш похилу. Таким чином, ефект угадування знижує диференціюючу здатність завдань тесту.

Оцінювання параметрів функції успіху

Під впливом випадкових факторів оцінки параметра δ, отримані на різних вибірках студентів, будуть відрізнятись. Якщо об'єм вибірки великий, то постає питання про обчислення стійких значень параметра δ, які будуть найбільш ефективними оцінками і можуть бути прийняті як об'єктивні значення параметра δ.

Одним із методів обчислення ефективних оцінок є метод найбільшої правдоподібності, запропонований Р. Фішером. Цей метод грунтується на використанні функції правдоподібності [6].

Нехай n учасників тестування виконують m різних завдань. $\{x_{ij}\}$ – матриця дихотомічних результатів, де $i = \overline{1,n}$, $j = \overline{1,m}$.

Функція правдоподібності L є добутком ймовірності $P_{ij} = \dfrac{e^{\theta_i - \delta_j}}{1 + e^{\theta_i - \delta_j}}$ для всіх можливих i та j, де $i = \overline{1,n}$, $j = \overline{1,m}$, θ_i – логіт рівня знань студента, δ_j – логіт складності завдань [8]:

$$L(x_{ij}, \theta_i, \delta_j) = \prod_{i=1}^{n}\prod_{j=1}^{m} P\{x_{ij} \mid \theta_i, \delta_j\} = \frac{\exp\left[\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}(\theta_i - \delta_j)\right]}{\left[\prod_{i=1}^{n}\prod_{j=1}^{m}(1 + \exp(\theta_i - \delta_j))\right]}.$$

Всі завдання тесту є локально незалежними. Це означає, що при даному рівні знань відповідь на кожне завдання тесту не залежить від результатів виконання решти завдань.

Як оцінку латентних параметрів приймають такі значення, при яких функція правдоподібності досягає максимуму. Такі оцінки називають оцінками найбільшої правдоподібності [11].

Оскільки функції $L(x_{ij}, \theta_i, \delta_j)$ та $\ln L(x_{ij}, \theta_i, \delta_j)$ досягають максимуму при одних і тих же значеннях, то зручніше шукати максимум логарифмічної функції правдоподібності $\ln L(x_{ij}, \theta_i, \delta_j)$.

$$\ln L(x_{ij}, \theta_i, \delta_j) = \sum_{i=1}^{n} b_i\theta_i - \sum_{j=1}^{m} c_j\delta_j - \sum_{i=1}^{n}\sum_{j=1}^{m} \ln\left[1 + \exp(\theta_i - \delta_j)\right],$$

де $b_i = \sum_{j=1}^{m} x_{ij}$ та $c_j = \sum_{i=1}^{n} x_{ij}$ – суми балів, набрані відповідно i-им студентом та за j-е завдання.

Невідомі оцінки найбільшої правдоподібності для параметрів θ і δ знаходимо з необхідної умови екстремуму логарифмічної функції правдоподібності по кожній із змінних θ_i та δ_j. Отже, для

знаходження максимуму функції правдоподібності прирівнюємо до нуля відповідні частинні похідні логарифмічної функції правдоподібності:

$$\frac{\partial \ln L(\theta_i, \delta_j)}{\partial \theta_i} = b_i - \sum_{j=1}^{m} \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)} = b_i - \sum_{j=1}^{m} P_{ij} = 0,\ i = \overline{1,n},$$

$$\frac{\partial \ln L(\theta_i, \delta_j)}{\partial \delta_j} = -c_j + \sum_{i=1}^{n} \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)} = -c_j + \sum_{i=1}^{n} P_{ij} = 0,\ j = \overline{1,m}.$$

Отримаємо систему із $n+m$ нелінійних рівнянь, що має єдиний розв'язок, який нелегко знайти. Тому проблемою даного методу є відшукання розв'язку даної системи.

Спочатку покладаємо відомі значення параметра δ_j, а θ_i розглядаємо як шукану змінну. Потім перевизначаємо значення θ_i, присвоюючи їм щойно знайдені значення, і шукаємо значення оцінки δ_j. Процес продовжуємо доти, поки модуль значення різниці в результаті ітерацій не стане меншим 0,01:

$$\left|\delta_j^{(k+1)} - \delta_j^{(k)}\right| < 0,01,$$

де k – кількість послідовних наближень.

Оцінки $\delta_j^{(k+1)}$ є найбільш ефективними і можуть бути прийняті за справжні значення латентних параметрів δ_j. Але для реалізації цього методу потрібні спеціальні програми.

Важливим моментом є правильний вибір гарного початкового наближення. Тому бажано при оцінюванні параметра δ використовувати формули, запропоновані Дж. Рашем для IRT [7, 8]. Хоча можливі й інші методи оцінювання початкових наближень. Якщо θ_i і δ_j вибрані невдало, достатньо далеко від оцінок найбільшої правдоподібності, то число ітерацій збільшиться. Відповідно зростуть витрати машинного часу.

5. Висновки

Описані основні класичні статистичні методи аналізу результатів тестування дозволяють провести найпростіші та необхідні процедури статистичної обробки результатів тестування знань і визначити методи оцінки якості тесту. Розглянутий підхід до конструювання тестів, згідно з яким рівень знань учасників тестування оцінюється за допомогою їх індивідуальних балів, а складність завдань – за допомогою частки правильних та неправильних відповідей на них, показав необхідність використовувати нові методи конструювання тестів, представлені в так званій сучасній теорії тестування на основі математичної теорії параметричної оцінки тестових завдань. У роботі розглянуто основні сучасні математичні моделі: однопараметрична модель Раша, двопараметрична та трипараметрична модель Бірнбаума, в яких параметри рівня підготовки θ та складності завдання δ виражаються як показники, задані в одній шкалі логітів, що дозволяє ввести взаємозв'язок між змінними у вигляді різниці $\theta - \delta$, коректно порівняти результати студентів,

отримані за допомогою різних тестів, оцінити рівень складності завдань незалежно від рівня підготовки груп студентів.

СПИСОК ЛІТЕРАТУРИ

1. Аванесов В.С. Теория и методика педагогических измерений (материалы публикаций). – М.: ЦТ и МКО УГТУ-УПИ, 2005. – 98 с.
2. Челышкова М.Б. Теория и практика конструирования педагогических тестов: Учебное пособие. – М.: Логос, 2002. – 432 с.
3. Люсин Д.В. Основы разработки и применения критериально-ориентированных педагогических тестов. – М.: Исследовательский центр, 1993. – 51 с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика: Учебное пособие для вузов. – 4-е изд., доп. – М.: Высшая школа, 1972. – 480 с.
5. Солодовников А.С. Теория вероятностей. – М.: Просвещение, 1978. – 192 с.
6. Аванесов В.С. Научные проблемы тестового контроля знаний. – М.: Учебный центр при ИЦПКПС, 1994. – 136 с.
7. Челышкова М.Б. Адаптивное тестирование в образовании (теория, методология, технология). – М.: Исследовательский центр проблем качества подготовки специалистов, 2001. – 165 с.
8. Челышкова М.Б. Разработка педагогических тестов на основе современных математических моделей: Учебное пособие. – М.: Исследовательский центр проблем качества подготовки специалистов, 1995. – 32 с.
9. Янченко С.И. Математическая модель оценки результатов тестирования // Тезисы докладов Всероссийской конференции «Развитие системы тестирования в России». – Москва, 2000.
10. Lord F.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale N-J. Lawrence Erlbaum Ass., 1980. – 266 p.
11. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. With a Foreword and Afteword by B.D. Wright. The Univ. of Chicago Press. – Chicago & London, 1980. – 199 p.

Стаття надійшла до редакції 14.04.2007