УДК 517.9

Глибовець М. М., д.ф.-м.н., проф..
Остапенко О. Ю., магістр НАУКМА

## Аналіз тестових завдань на основі статистичної обробки результатів тестування

*Розгядаються типи тестів, які можуть бути використано в дистанційній освіті. Наведено підхід до статистичної обробки результатів тестів, а також описано основні характеристики тестів, такі як надійність та валідність.*

*Ключові слова: характеристики тестів, валідність, надійність.*

Glybovets M. M., DPhil
Ostapenko O. Y., Master of computer science.

## Tests analysis based on the static processing of their results.

*Given article describes different test types such as adaptive testing that could be utilized in distance learning. You can find an approach to perform static analysis of the test results along with the description of the main test characteristics – validity and reliability.*

*Key words: test characteristics, validity, reliability.*

*E-mail: oleksandra@gmail.com
Статтю представив д.ф.-м.н., проф. Анісімов А.В.

**Вступ** Контроль якості навчальної роботи є важливим засобом управління процесом навчання, особливо це стосується систем дистанційного електронного навчання, які базуються на самостійній роботі слухача курсу. Необхідність контролю навчальної роботи й оцінки знань має об'єктивний характер. Тут діє закономірний зв'язок у ланцюгу: мета навчання – процес – результат – нова мета. Для визначення нової мети, необхідно точно знати, що вже досягнуто внаслідок навчання.

Тестування – один з найпоширеніших методів контролю в сучасній системі освіти.

Комплексне і об'єктивне оцінювання знань через надання ефективного зворотного зв'язку є запорукою успішного навчання. Різноманітність систем тестування вражає. Вони різняться за багатьма критеріями: формами тестових запитань, видами тестів, взаємодією з користувачем, методами оцінювання ефективності результатів і параметрами інтерфейсу. Створена для конкретного навчального курсу або конкретної предметної області окрема підсистема тестування має тільки специфічний набір функцій та параметрів, що допомагає вирішити конкретну проблему. Це призвело до розробки у світі величезного різноманіття вузькоспеціалізованих систем тестування, несумісних між собою, з мінімальною специфічною функціональністю. Практично кожна система автоматизованого навчання включає в себе свій варіант підсистеми тестування. Такі розробки дуже схожі між собою, причому більшість з них реалізують тільки найлегші сценарії тестування, залишаючи комплексний математичний підхід до оцінки знань на майбутнє.

Важливою задачею є вибір технології та архітектурного підходу, який би дозволив створити гнучку платформу побудови систем тестування і легкого їх налаштування на будь-який навчальний курс будь-якої навчальної платформи без обмеження на види тестування чи на типи тестових питань, формат тестів. Така платформа повинна мати чітко визначений опис інтерфейсів та основних компонентів системи, для подальшої тісної інтеграції з будь-якими системами навчання без необхідності вносити зміни в вихідні коди системи тестування.

Окрім цього, важливою також є можливість автоматизації побудови тестових завдань з орієнтацією на побудову валідних, надійних і функціонально повних тестів [1]. Для визначення цих характеристик застосовується певна статистична обробка матеріалів навчання. В даній статті буде розглянуто основні відомі підходи для такого аналізу. Основний акцент зробимо на ймовірносних характеристиках. Аналіз завдань тестування математичними методами дозволяє одержати інформацію про їх приховані дефекти, що не можуть бути виявлені за допомогою експертних методів.

**Статистична обробка результатів тестування**

Розглянемо найпростіші і необхідні процедури статистичної обробки результатів тестування знань і методи оцінки якості тесту відповідно до класичної теорії тестування [2].

Позначимо через $x_{ij}$ числову оцінку успішності виконання j-го завдання, виконаного і-м студентом. Результати тестування звичайно представляються у вигляді матриці $\{x_{ij}\}$ з n рядками і m стовпцями *(i=1,…,n; j=1,…,m)*. У практиці тестування прийнято, як правило, користуватися дихотомічною шкалою оцінок результатів, коли множина можливих оцінок складається усього з двох елементів {0;1}: 0 - завдання не виконане, 1 - виконане правильно. Це, звичайно, не єдина можлива шкала. Розрахунок, однак, виконується по формулах, що наведено нижче, незалежно від обраної для оцінок шкали.

Процес статистичної обробки матриці результатів тестування будемо розглядати послідовно, по кроках.

• **1 крок.** Обчислюються індивідуальні бали студентів $y_i$ (*i=1,…,n*), що показують результат проходження тесту кожним з них:

$$y_i = \sum_{j=1}^{m} x_{ij}.$$

Оскільки для перевірки статистичних гіпотез, що застосовуються в класичній теорії тестів, використовують припущення про нормальний розподіл сумарних балів студентів, то рекомендується досліджувати розподіл частот. Для порівняння розподілу балів з нормальним можна використовувати будь-який з критеріїв, що звичайно застосовуються для цієї мети.

• **2 крок.** Обчислюються середні результати $\bar{y}$ сумарних балів студентів та середні результати $\bar{x}_j$ студентів по кожному завданню.

Для дихотомічних даних дані, що обчислюються по аналогічній формулі, позначаються через $p_j$ і традиційно називаються в тестології мірою складності завдання *j (j=1,2,…,m):*

$$p_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}.$$

Відмітимо, однак, що чим більша величина коефіцієнта $p_j$, тим більша частина студентів успішно впорюється з завданням *j*. Так що насправді коефіцієнти $p_j$ *(j=1,2,...,m)* повинні інтерпретуватися як показники легкості завдань.

• **3 крок.** Обчислюється дисперсія $s_y^2$ і стандартне відхилення $s_y$ сумарних балів студентів:

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

• **4 крок.** Обчислюється дисперсія $s_y^2$ результатів студентів по *j*–ому завданню *(j=1,…,m)*. Якщо успішність виконання завдання оцінюється балами 0 або 1, міра варіації визначається по формулі: $s_j^2 = p_j \times \left(1 - p_j\right)$.

Коли множина оцінок складається з більш ніж двох значень, то використовується формула:

$$s_j^2 = \frac{\sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2}{n-1}.$$

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

• **5 крок.** Визначається зв'язок кожного *j*–го завдання (*j=1,…,m)* з сумою балів по всьому тесту за допомогою коефіцієнту кореляції Пірсона:

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

• **6 крок.** Визначається попарний кореляційний зв'язок завдань між собою. Тут теж можна використовувати коефіцієнт кореляції Пірсона.

• **7 крок.** Обчислюється індекс $I_j$(*j=1, 2, … m*) дискримінативності завдання, тобто його розрізняюча властивість, що вказує на можливість розділяти окремих студентів за рівнем виконання тесту в цілому. Для цього з загальної сукупності студентів виділяють дві підгрупи - тих, хто одержав найвищі сумарні бали, і тих, хто одержав найнижчі . Тоді індекс дискримінативності може бути визначений як різниця між відносними кількостями студентів, що правильно виконали завдання j в цих двох підгрупах. Наприклад, впорядковану сукупність сумарних балів поділяють на три частини і порівнюють результати виконання кожного завдання j першою й останньою третинами студентів. У цьому випадку для дихотомічних даних індекс здобуває вигляд:

$$I_j = \frac{\sum_{i=1}^{n/3} x_{ij} - \sum_{i=2n/3+1}^{n} x_{ij}}{n/3}$$

Чим більше коефіцієнт $I_j$, тим більше дискримінативність завдання.

• **8 крок.** Черговий крок робиться на основі вектора кореляцій $\{R_j\}$ (або $\{B_j\}$), кореляційної матриці $\{\varphi_{jk}\}$ (або $\{r_{jk}\}$) і вектора коефіцієнтів складності $\{p_j\}$. Із сукупності тестових завдань видаляються завдання, що не мають дискримінативності, тобто занадто легкі завдання ($p_j$>0,9) і занадто складні ($p_j$ <0,2). Після цього виключаються завдання, що погано корелюють з сумую балів ($R_j$<0,15), і які мають від'ємні коефіцієнти кореляції $\varphi_{jk}$ (або $r_{jk}$).

• **9 крок.** Для скороченого переліку завдань знову обчислюються сумарні бали студентів $y_i$. Потім складається нова, впорядкована, матриця даних тестування, у якій стовпці розташовуються в порядку зростання складності завдань, а рядка - у порядку зменшення, зверху вниз, сумарних балів студентів. Для зредукованої матриці перераховуються середній сумарний бал, дисперсія сумарних балів і коефіцієнти кореляції завдань із сумою балів [3].

Застосувавши даний алгоритм до існуючих тестів, викладач може з'ясувати наскільки реальні результати студентів відповідають бажаним результатам, і, у разі потреби, підкоригувати тест під рівень студентів або під рівень усієї групи.

Було проведено декілька експериментів, в результаті яких ми дійшли висновку, що статистичні методи дослідження не завжди застосовні через їх громіздкість і достатню складність. Крім того, необхідність в обчисленнях такого роду виникає не досить часто. Але незважаючи на це все одно бажано робити перевірку тестів для того, щоб вони якомога більше відповідали лекційним матеріалам та дозволяли вірно оцінити знання студентів. Як уже було сказано вище, головними характеристиками тесту є надійність і валідність. Їх теж можна обрахувати за допомогою математичних моделей.

**Надійність**

Надійність тесту ρ тим вище, чим більш узгоджені результати того самого студента при повторній перевірці знань за допомогою того ж тесту або еквівалентної його форми (паралельного тесту) [4]. Узгодженість результатів можна вимірювати коефіцієнтом кореляції Пірсона. Якщо значення коефіцієнта попадають в інтервал 0,80-0,89, то говорять, що тест має гарну надійність, а якщо цей коефіцієнт не менше 0,90, то надійність можна назвати дуже високої.

Інші, більш практичні, методи оцінки надійності тесту, засновані на однократному застосуванні єдиної форми тесту [5].

Оцінку надійності повного тесту можна робити з використанням коефіцієнта кореляції $r_{1/2}$, по формулі Спірмана-Брауна:

$$\rho = \frac{2r_{1/2}}{1 + r_{1/2}}.$$

Інший спосіб оцінки надійності розщепленого тесту заснований на формулі Рюлона:

$$\rho = 1 - \frac{s_d^2}{s_y^2},$$

$s_y^2$ - дисперсія сумарних балів результату, а $s_d^2$ - дисперсія різниць між результатами кожного студента по обох половинах тесту. Вона обчислюється по формулі:

У тестуванні цей метод визначення надійності підходить для порівняно гомогенних за формою тестів, що містять завдання приблизно одного рівня складності. Припустимо розглядати внесок окремого тестового завдання в надійність усього тесту і узгодженість результатів відповідей на дане завдання з результатами усього тесту. Відповідно до цього було виведено індекс надійності окремого тестового завдання.

**Валідність**

Валідність тесту показує, наскільки добре тест робить те, для чого він був створений, тобто це комплексна характеристика тесту, що відображає обґрунтованість, значимість результатів, адекватність тесту цілям виміру.

Визначити коефіцієнт валідності тесту - означає визначити, як виконання тесту співвідноситься з іншими незалежно зробленими оцінками знань студентів. Для визначення валідності потрібен незалежний зовнішній критерій, тобто оцінка експерта (викладача). За коефіцієнт валідності приймають коефіцієнт кореляції результатів тестових вимірів і критерію. Якщо експертна оцінка знань студентів, що отримана незалежно від процедури тестування, представлена числовою

послідовністю $Y_1, Y_2, \ldots, Y_n$, то коефіцієнт валідності тесту може бути обчислений за формулою:

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

де $\overline{Y}$ - середнє арифметичне експертних оцінок, $s_Y$ - стандартне відхилення цих оцінок.

Виходячи з того, що валідність визначається на підставі порівняння результатів тестування і показників, отриманих незалежним шляхом (традиційних оцінок, експертних суджень, результатів інших тестів, валідність яких було встановлено попередньо) і звичайно обчислюється за допомогою коефіцієнтів кореляції, то з двох тестів, призначених для однієї і тієї ж мети, більш *ефективним* буде той, який швидше, дешевше і якісніше вимірює знання даної групи студентів. Під швидкодією в даному випадку розуміємо таку кількість питань яка дозволяє отримати вірне представлення про рівень знань студента.

Зрозуміло, що чим вища валідність, тим більш обґрунтоване використання результатів тестування для висновків, аналізу і передбачень.

Відмітимо, що валідність тестів може поділятися на декілька видів. Наприклад, розрізняють критеріальну валідність (Criterion validity), змістовну (Content Validity) і конструктивну(Construct Validity) [6].

Критеріальна валідність тесту – це характеристика, що відображає ступінь впевненості, що отримані оцінки реально відображують досягнення визначеного рівня знань або навичок.

Змістовна валідність – це характеристика тесту, що відображує ступінь впевненості, що тестові завдання досить повно відображують зміст визначеної області знань (предметної області), а володіння навичками, що перевіряються в тесті, важливе для даної діяльності, і при цьому тест не перевіряє другорядних або непотрібних знань та навичок.

Конструктивна валідність тесту – це характеристика, що відображує впевненість в тому, що властивість, що вимірює тест, являє собою конструкт, що займає певне місце в теорії даної області знань.

**Типи тестів**

Тести можуть розрізнятися за орієнтованістю та за типом генерації питань. За орієнтованістю розрізняють крітеріально-орієнтований тест та нормативно-орієнтовний [7]. Крітеріально-орієнтованим вважають тест що призначений для виміру тієї частини учбового матеріалу, яка засвоєна студентом, або для визначення відповідності студента визначеному критерію. Крітеріально-орієнтовані тести протиставляються нормативно-орієнтованим тестам, які оцінюють досягнення даного студента у порівнянні з досягненнями інших. Обидва типи можуть включати ті самі (однакові) завдання, різниця полягає у інтерпретації результатів.

За типом генерації питань розрізняють 2 типи тестів – тести, що є завжди сталими, та тести, що генеруються випадковим чином з бази питань. Сталий тест (для нього також використовується назва «авторський») представляє собою набір чітко визначених питань, які постійні для кожного тесту. Кожен студент отримує даний тест без змін. Цей тест корисний для використання на екзаменах, а також в процесі засвоєння знань для самоконтролю. Такі тести дають можливість побачити рівень успішності групи в цілому.

Тести, що генеруються випадковим чином (також мають назву «тематичні») представляють собою динамічно змінні тести, при створенні яких задається тематика цього тесту, а також кількість питань з кожної теми. Тест генерується автоматично на основі цих даних, вибираючи з банку даних питань питання випадковим чином.

Цей тип тесту при умові великого банку даних питань гарантує генерацію практично повністю різних тестів. Цей тип зручно застосовувати для контрольних тестів. Такий тест дає можливість автоматично побудувати індивідуальні тести для кожного студента та перевірити його власний рівень знань.

Різновидом такого типу тестів є адаптивний тест, тобто тест, завдання якого пред'являються студенту в залежності від того, як він виконав попереднє завдання або сукупність попередніх завдань [7]. Загальне правило адаптивного тесту: ефективне виконання завдання є підставою для пред'явлення наступного завдання більших труднощів, а неефективне виконання завдання - підставою для пред'явлення менш важкого завдання, тобто порядок питань може бути не тільки лінійним, але й залежати від попередніх відповідей студента.

Таким чином, адаптивний тест являє собою варіант автоматизованої системи тестування з заздалегідь відомими параметрами складності.

Розрізняють 3 типи адаптивного тесту:

- Перший називається *пірамідальним тестуванням*. При відсутності попередніх оцінок усім студентам дається завдання середньо складності і вже потім, в залежності від відповіді, кожному випробуваному дається завдання легше або складніше; на кожнім кроці корисно використовувати правило розподілу шкали складності навпіл.
- Другий варіант - *flexilevel*- контроль починається з будь-якого рівня складності, з поступовим наближенням до реального рівня знань.
- Третій варіант - *stradaptive* (від англ. stratified adaptive), коли тестування проводиться за допомогою банку завдань, розділених по рівнях складності. При правильній відповіді наступне завдання береться з верхнього рівня, при неправильному - з нижнього.

Крім вищезазначених типів тестів розрізняють ще *гомогенні*, *гетерогенні* та *інтегративні*.

Гомогенний тест являє собою систему завдань зростаючої складності, специфічної форми і визначеного змісту - систему, створену з метою об'єктивного, якісного, і ефективного методу оцінки структури і виміру рівня підготовленості студентів по одній навчальній дисципліні.

Гетерогенний тест являє собою систему завдань зростаючої складності, специфічної форми і визначеного змісту - систему, створену з метою об'єктивного, якісного, і ефективного методу оцінки структури і виміру рівня підготовленості студентів по декількох навчальних дисциплінах.

Інтегративним можна назвати тест, що складається із системи завдань, що відповідають вимогам інтегративного змісту, тестової форми, зростаючої складності завдань, націлених на узагальнену підсумкову діагностику підготовленості студента освітнього закладу.

Перевага інтегративних тестів перед гетерогенними полягає в більшій змістовній інформативності кожного завдання й у меншому числі самих завдань.

**Список використаних джерел**

1. *Аванесов В.С.* Научные основы тестового контроля знаний. -М.: Исследовательский центр, 1994. – 135 с.
2. *Аванесов В.С.* Математические модели педагогического измерения. -М.: Исследовательский центр проблем измерения качества подготовки специалистов, 1994. - 26с.
3. *Lord F.M.* Application of Item Response Theory to Practical Testing. – N - Y., Academic Press, 2001 – 250с.
4. *Keeves J.P.* Educational Research, Methodology and Measurement: An International Handbook. –Oxford, Pergamon Press, 1988.
5. *Челишкова М. Б.* Теория и практика конструирования педагогических тестов. Учебное пособие. – М.: Логос, 2002, – 432 с.
6. *Weiss D.J.(Ed.)* New Horizons in Testing: Latent Trait Test Theory and Computerised Adaptive Testing. – N-Y., Academic Press, 1983.-345с.
7. *Гласс Д., Стенли Д.* Статистические методы в психологии и педагогике – М.: 1981.-140с.

Надійшла до редколегії 11.12.2009