Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
https://nv.nltu.edu.ua
https://doi.org/10.15421/40280824
Article received 11.10.2018 p.
Article accepted 25.10.2018 p.
УДК 004:519.25

ISSN 1994-7836 (print)
ISSN 2519-2477 (online)

Correspondence author
Ya. F. Kuleshnyk
kuleschnyk@gmail.com

Я. Ф. Кулешник, В. В. Сеник, Т. В. Магеровська
Львівський державний університет внутрішніх справ, м. Львів, Україна

ВИКОРИСТАННЯ СТАТИСТИЧНИХ МЕТОДІВ ДЛЯ СТВОРЕННЯ КОГНІТИВНИХ ТЕСТОВИХ ЗАВДАНЬ

Подано означення тесту, складності тестового завдання, валідності, дискримінативності та надійності тестових завдань. Обґрунтовано необхідність проходження тестовими завданнями процесу спеціального оцінювання з використанням елементів математичної статистики. Розглянуто два основні підходи до створення тестів та головне питання теорії тестів – побудова оптимального тесту. Подано математичну постановку задачі створення і аналізу тестових завдань. На конкретних результатах тестових завдань розглянуто методику розрахунку тестових характеристик. Складність тестових завдань розкрито на підстаі емпіричної перевірки завдань, з підрахунком частки правильних відповідей, за відомими формулами розраховано коефіцієнти варіації та дисперсії. Зроблено оцінку валідності з використанням значущого зовнішнього критерію – експертної оцінки. Подано таблицю критеріїв для інтерпретації значення коефіцієнта валідності, щоб унеможливити створення неякісних тестових завдань. Для обчислення коефіцієнта дискримінативності застосовано метод крайніх груп, подано критерії інтерпретації значення коефіцієнта дискримінативності, розроблено конкретні рекомендації для створення тестових завдань з належним значенням коефіцієнта дискримінативності. Розглянуто два види надійності тестів: надійність як стійкість і надійність як внутрішню погодженість. Для визначення значення надійності як стійкості застосовано перетворений коефіцієнт кореляції Пірсона "*фі*" для дихотомічних даних. Наведено результати проведених розрахунків значення коефіцієнта надійності як внутрішньої погодженості вирахуваного за формулою Спірмена-Брауна з використанням формули Пірсона. Подано таблицю критеріїв для інтерпретації значення коефіцієнта надійності, використано метод Кьюдера-Річардсона як метод розщеплення тесту для оцінки надійності дихотомічних завдань тестів. Наведено інтерпретацію результатів та вироблення спрощених загальних рекомендацій укладачам, котрі можна застосовувати в тестологічній практиці.

Ключові слова: адаптивний тест; класичні статистичні методи; складність; валідність; дискримінативність; надійність тестових завдань.

Вступ. Тест – це інструмент, що складається з вивіреної системи завдань, стандартизованої процедури проведення і наперед спроектованої технології аналізу результатів для визначення рівня знань, умінь, навиків, властивостей характеру особистості, розумових здібностей, зміна яких можлива у процесі систематичного навчання (Avanesov, 2005).

Застосування тестового контролю під час проведення вступних компаній у ВНЗ, перевірення залишкових знань учнів шкіл, ліцеїв, коледжів та університетів – одне з актуальних завдань сьогодення (Chelyshkova, 2002).

Головна проблема тестового контролю знань – сам процес створення тестів, їхня уніфікація та методи проведення аналізу. Щоб довести тест до повної готовності, для використання необхідна багаторічна кваліфікована праця із складання позбавлених суб'єктивізму у формулюванні тестових завдань та збору статистичних даних. Для об'єктивної оцінки рівня знань необхідне професійне і грамотне складання тесту. Недостатньо придумати запитання і варіанти відповідей, оскільки в цьому випадку може виникнути немало суперечностей, помилок, невизначеності, завдання можуть бути дуже простими або ж, навпаки, надто складними. Саме тому тестові завдання повинні проходити процес спеціального оцінювання з використанням елементів математичної статистики. Одне з головних питань теорії тестів – побудова оптимального тесту.

Історично виділяють два основні підходи до створення тестів. Перший з них набув значного розвитку в рамках класичної теорії тестів. Згідно з ним, рівень знань учасників тестування оцінюють за допомогою балів, набраних під час тестування. Бал обчислюють як алгебраїчну суму оцінок виконання кожного завдання тесту. Статистичні методи аналізів результатів лягли в основу класичної теорії тестування знань і методів

Інформація про авторів:

Кулешник Ярко Федорович, канд. техн. наук, доцент, кафедра інформатики. **Email:** kuleschnyk@gmail.com; https://orcid.org/0000-0002-0707-9087

Сеник Володимир Васильович, канд. техн. наук, доцент, завідувач кафедри інформатики. **Email:** v.v.senyk@gmail.com; https://orcid.org/0000-0002-0428-6443

Магеровська Тетяна Валеріївна, канд. фіз.-мат. наук, доцент, кафедра інформатики. **Email:** magerovskat@gmail.com; https://orcid.org/0000-0001-6763-4321

Цитування за ДСТУ: Кулешник Я. Ф., Сеник В. В., Магеровська Т. В. Використання статистичних методів для створення когнітивних тестових завдань. Науковий вісник НЛТУ України. 2018, т. 28, № 8. С. 122–128.

Citation APA: Kuleshnyk, Ya. F., Senyk, V. V., & Maherovska, T. V. (2018). The use of statistical methods for creation of cognitive test tasks. *Scientific Bulletin of UNFU, 28*(8), 122–128. https://doi.org/10.15421/40280824

оцінки якості тестів.

Метою роботи є часткова систематизація найпростіших методів, що дають змогу розраховувати тестові характеристики, а саме: визначення валідності, дискримінативності та надійності деяких тестів, що проводились у ЛьвДУВС, інтерпретація результатів та вироблення спрощених загальних рекомендацій укладачам.

Виклад основного матеріалу. Математична постановка задачі створення тестових завдань. Позначимо через $x_{i,j}$ числову оцінку успішності виконання j-го завдання i-им студентом. Результат тестування звичайно можна представити у вигляді матриці $\{X_{i,j}\}$, що містить n рядків та m стовпців $(i=\overline{1,n}; j=\overline{1,m})$. Ця матриця показує результат виконання всіх завдань усіма учасниками тестування. На практиці зазвичай використовують дихотомічну шкалу оцінок результатів, тобто внаслідок правильного виконання завдання тестований отримує один бал, $x_{i,j}$ =1, у протилежному випадку – нуль балів, $x_{i,j}$ = 0. У такому випадку результатом виконання тесту буде кількість правильних відповідей. Результат можна оцінювати не тільки нулем чи одиницею, але й присвоювати певний ваговий коефіцієнт, що відповідає складності завдання, однак істотних змін у якість оцінювання тестових завдань він не дає (Bovtrukevich & Kireenko, 2017).

Для майбутніх обчислень індивідуальний початковий бал (результат, або кількість правильних відповідей за всі завдання) i-го тестованого після проходження тесту позначимо y_i $(i=\overline{1,n})$, середній результат сумарних балів усіх учасників тестування – $\overline{y}$, середній результат тестованого за кожним завданням $\overline{x}_j, (j=\overline{1,m})$:

$$y_i=\sum_{j=1}^{m} x_{i,j}, i=\overline{1,n};\ \overline{x}_j=\frac{1}{n}\sum_{i=1}^{n} x_{i,j}, j=\overline{1,m};\ \overline{y}=\frac{1}{n}\sum_{i=1}^{n} y_i . \quad (1)$$

Важливою вимогою до тестових завдань є їх об'єктивний рівень складності. Не можуть у тесті бути завдання з невідомою мірою складності. Завдання можна включати в тест тільки після емпіричної перевірки їх міри складності.

Методику розрахунку тестових характеристик покажемо на конкретному прикладі. У тестуванні брали участь 29 студентів. Тест складався з 30 запитань. До кожного запитання надавалось 3–4 відповіді, серед яких потрібно було обрати правильну. Правильну відповідь оцінювали 1 балом, неправильну – 0 балів. Ми отримали результати тестування студентів, що показані в табл. 1.

Складність тестового завдання. Складність завдання можна визначити двома способами (Bovtrukevich & Kireenko, 2017):

- на основі оцінки передбачуваного числа і характеру розумових операцій, необхідних для його вдалого виконання;
- на основі емпіричної перевірки завдань, з підрахунком частки правильних відповідей.

Протягом багатьох років у класичній теорії тестів розглядали тільки емпіричні показники складності. На сьогодні в дистанційному навчанні застосовують сучасні теорії навчальних тестів, де більше уваги приділяють характеру розумової діяльності у процесі виконання тестових завдань різних форм.

Емпірично складність завдання визначають додаванням елементів матриці $\{X_{i,j}\}$ по стовпцях, що дорівнює числу правильних відповідей, отриманих за кожним стовпцем (R_j). Чим більше правильних відповідей на конкретне завдання, тим воно легше для відповідної групи студентів. З огляду на різну кількість тестованих у вибірці для одержання об'єктивних характеристик R_j ділять на кількість в кожній групі n.

$$p_j=\frac{R_j}{n}, j=\overline{1,m} . \quad (2)$$

Внаслідок отримаємо нормований статистичний показник – частка правильних відповідей, p_j. Показник p_j тривалість використовували як показник рівня складності завдання у класичній теорії тестів. Пізніше усвідомили, що зі збільшенням значення p_j складність не зростає, а навпаки – знижується. Тому ввели показник складності, що відображав відношення кількості неправильних відповідей W_j до кількості учасників тестування n:

$$q_j=\frac{W_j}{n},\ p_j+q_j=1, j=\overline{1,m} . \quad (3)$$

Завдання є не тестовим, якщо на нього правильно чи неправильно, тобто однаково, відповідають усі тестовані. Між ними відсутня варіація, іншими словами – варіація дорівнює нулю. Нульова варіація практично означає необхідність викидання завдання з тесту.

Зручною мірою варіації може бути значення дисперсії s_y^2, чи стандартне відхилення s_y сумарних балів кожного тестованого та величина s_j^2 – дисперсії результатів тестованих по j-му завданню:

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

$$s_j^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j}-x_j)^2, j=\overline{1,m} . \quad (4)$$

У випадку, коли результат виконання j-го завдання оцінюють балами 0 чи 1, то міру варіації визначають за формулою $s_j^2=p_j\cdot(1-p_j)$ або $s_j^2=p_j\cdot q_j, j=\overline{1,m}$.

Результати розрахунків наведено в табл. 1.

Валідність тестів. Подано декілька визначень поняття валідності (Bovtrukevich & Kireenko, 2017):

1. Валідність – придатність тестових результатів для тієї мети, заради чого проводилось тестування.
2. Валідність – це характеристика вміння тесту служити поставленій меті вимірювання.
3. Валідність – визначає, наскільки тест відображає те, що він повинен оцінювати.

Для оцінки валідності тесту зазвичай використовують кореляцію між показниками тесту і деяким зовнішнім критерієм. За такої оцінки дуже важливо вибрати значущий зовнішній критерій. Процес валідації у цьому випадку ускладнюються необхідністю встановлення міри узгодженості оцінок експертів, котрих зазвичай буває не менше трьох чоловік. Для педагогічних тестів у якості критерію звичайно беруться оцінки, що були виставлені під час традиційного перевірення знань студентів без застосування тестів, або на підстаі поточної успішності.

Розрізняють різноманітні види валідності: змістовна, концептуальна, критеріальна, поточна, прогностична та ін.

Валідність визначають за формулою:

$$V=\frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i\cdot y_i)-\overline{Y}\cdot\overline{y}}{S_Y-S_y}\cdot\frac{n}{n-1}, \quad (5)$$

де: n – кількість студентів; i – порядковий номер студента; Y_i – експертна оцінка i-го студента; $\overline{Y}$ – середнє арифметичне експертних оцінок; S_y – стандартне відхилення кількості правильних відповідей; S_Y – стандартне відхилення експертних оцінок, визначають як

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i\,,\ S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}\,,\ S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}\,. \quad (6)$$

Табл. 1. Розрахунок коефіцієнта валідності тестового завдання

Номер запитання									
Кандидат	1	2	3	…	29	30	Разом	Експертна оцінка	Стандартне відхилення результату (S_y)
1	0	0	0	…	0	0	5	6	5,5
2	0	0	1	…	1	0	13	12	Стандартне відхилення експертних оцінок (S_Y)
3	0	0	1	…	1	0	7	8	4,6
4	1	0	0	…	0	0	10	11	
5	0	0	0	…	1	1	8	7	
…	…	…	…	…	…	…	…	…	Валідність (V)
26	0	0	0	…	1	1	6	7	0,97
27	0	0	0	…	1	0	7	8	
28	0	0	0	…	0	0	6	6	Середнє значення дисперсії (S_i^2)
29	0	0	1	…	0	0	10	12	0,2
Всього правильних відповідей	10	6	8	…	18	7	310	304	
Частка правильних відповідей	0,34	0,21	0,28	…	0,62	0,24			
Частка неправильних відповідей	0,66	0,79	0,72	…	0,38	0,76			
Дисперсія (S_i^2)	0,23	0,16	0,20	…	0,24	0,18			
Середній бал							10,69	10,48	
Maxs(i)	29	29	29	…	29	29	870		
P-показник складності	0,34	0,21	0,28	…	0,62	0,24	0,36		

Проведемо розрахунки, враховуючи формули (5), (6) та табл. 1, і отримаємо значення коефіцієнта валідності. У цьому випадку значення коефіцієнта валідності дорівнює 0,97, що свідчить про високу валідність тестових завдань, тобто тест достатній для того, щоб прийняти рішення хто як вчиться. Для інтерпретації значення коефіцієнта валідності застосовують такі критерії:

Значення коефіцієнта	Інтерпретація
від 0,6 до 1,0	висока валідність тесту
від 0,3 до 0,6	середня валідність тесту
< 0,3	низька валідність тесту

Для того, щоб підвищити валідність тесту, завдання повинні мати оптимальну складність. Щоб забезпечити нормальний закон розподілу балів по тесту, необхідно провести експертизу якості змісту тесту за допомогою третіх осіб, час виконання тесту повинен бути оптимальним, завдання повинні бути з високою дискримінативністю.

Дискримінативність тестів. Дискримінативність – це здатність тестових завдань диференціювати (розділити) студентів за ознакою максимального чи мінімального результату. Це поняття введене для того, щоб унеможливити створення неякісних тестових завдань.

Для обчислення коефіцієнта дискримінативності застосуємо метод крайніх груп. Суть цього методу полягає в тому, що під час розрахунку дискримінативності тестового завдання враховують результати найбільш та найменш успішних тестованих осіб. Кількість учасників крайніх груп залежить від величини вибірки, а саме: чим більша вибірка, тим меншу кількість тих, що підлягають тестуванню, можна залучати в обидві групи. Кількість тестованих у кожній групі повинна бути однаковою і знаходитись в межах від 10 % до 33 % від загальної кількості тестованих. У цьому випадку використаємо 27 % для кожної групи, тобто по 7 осіб у групі, тому що ця частка у теорії вважають найкращим для забезпечення максимальної точності визначення коефіцієнта дискримінативності. Індекс дискримінативності обчислюють як різницю між частиною осіб, що правильно розв'язали задачу, з найбільш і найменш успішної групи.

$$D = \frac{Nn_{\max}}{N_{\max}} - \frac{Nn_{\min}}{N_{\min}}\,, \quad (7)$$

де: N_{min}=N_{max} – загальна кількість тестованих у крайніх групах (27 %); Nn_{min} – кількість студентів у групі гірших, що правильно виконали завдання тесту; Nn_{max} – кількість студентів у групі кращих, що правильно виконали завдання тесту. Результати тестування подано у табл. 2. У цьому випадку значення коефіцієнта дискримінативності дорівнює 0,39, що свідчить про високу дискримінативність тестових завдань.

Коефіцієнт дискримінативності може приймати значення у межах від –1 до +1. Високе позитивне значення коефіцієнта дискримінативності тестового завдання свідчить про правильний розподіл тестованих на групи. Високе від'ємне значення свідчить про непридатність цієї задачі для цього тесту, тобто її невідповідність сумарному результату. У нашому випадку з табл. 2 видно, що завдання 5, 28, 29, 30 – неякісні, краща група відповідає гірше, ніж слабша.

Для забезпечення високого рівня дискримінативності потрібно, щоб ваші тести не були занадто складними; формулювання повинні бути прозорі; жодних неоднозначностей; розв'язки повинні бути неочевидними; варіанти відповідей повинні бути реальні, тобто не абсурдні; не може бути декілька варіантів відповідей, якщо вони не обговорені спеціально.

Надійність тестів. Надійністю називають характеристику тесту, що відображає точність тестових замірів, а також стійкість тестових результатів до дії зовнішніх чинників. Тест вважають надійним тільки у тому випадку, коли він забезпечує високу точність вимірювання та близькі результати внаслідок повторного

тестування цих же осіб за короткий проміжок часу. Тобто можна вважати, що надійність тесту показує, на скільки точно цей тест, як інструмент вимірювання, визначає знання студента чи якогось іншого явища. Розрізняють два види надійності тестів: надійність як стійкість і надійність як внутрішня погодженість.

Табл. 2. Розрахунок коефіцієнта дискримінативності тестового завдання

Кандидат	1	2	3	4	5	…	27	28	29	30	Разом
19	0	0	0	0	0		1	0	1	0	11
6	1	0	0	0	1		1	0	0	0	11
2	0	0	1	0	0		1	0	1	0	12
7	1	1	1	1	0		1	1	0	0	19
24	1	0	1	1	1		1	0	1	1	21
16	1	1	1	0	1		1	0	1	1	23
17	1	1	1	1	0		1	1	1	1	27
Nmax	5	3	5	3	3		7	2	5	3	
1	0	0	0	0	1		0	0	0	0	5
25	0	0	0	0	0		0	0	0	0	5
26	0	0	0	0	0		0	1	1	1	6
28	0	0	0	0	0		0	0	0	0	6
3	0	0	1	0	0		0	0	1	0	7
27	0	0	0	0	0		0	0	1	0	6
5	0	0	0	0	1		0	0	1	1	7
Nmin	0	0	1	0	2		0	1	4	2	
Індекс дискримінативності	0,7	0,4	0,6	0,4	0,1	…	1,0	0,1	0,1	0,1	0,39

Проведемо розрахунки, враховуючи формулу (7), і отримаємо значення коефіцієнта дискримінативності. Для інтерпретації значення коефіцієнта дискримінативності застосовують такі критерії:

Значення коефіцієнта	Інтерпретація
від 0,3 до 1,0	висока дискримінативність тесту
від 0,1 до 0,3	завдання потрібно проаналізувати на придатність до використання в тесті (низька диференціальна здатність)
< 0,1	завдання неякісне – краща група відповідає гірше, ніж слабша

Надійність як стійкість вимірюється за допомогою повторного тестування на тих же тестованих осіб не пізніше двох тижнів після першого тестування. Це значно ускладнює визначення надійності як стійкості тесту, бо пов'язане з додатковими організаційними труднощами. Відсутність декількох тестованих осіб, особливо за невеликих вибірок, під час повторного тестування, дає підстави сумніватися у надійності результатів тестування. Однак, хто дасть гарантії, що тестовані додатково не займалися протягом двох тижнів для покращення результатів.

За потреби для знаходження цієї характеристики можна застосовувати формулу Пірсона:

$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n} (X_i \cdot Y_i) - \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Y_i}{\sqrt{\left[n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2 \right] \cdot \left[n \cdot \sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i \right)^2 \right]}}, \qquad (8)$$

де: X_i – тестовий бал i-го тестованого під час першого вимірювання; Y_i – тестовий бал i-го тестованого під час повторного вимірювання.

Досліджуючи зв'язок між наборами даних, необхідно правильно вибирати вид і форму показника, що залежать від шкали, в якій представлені дані. Зокрема, для оцінювання зв'язку між результатами виконання тетстованими двох завдань тесту, коефіцієнт кореляції Пірсона r необхідно перетворювати, тому що результати виконання завдань представляють у дихотомічній шкалі (стовпці та рядки матриці представлені у вигляді нулів та одиниць). Перетворений коефіцієнт кореляції Пірсона для дихотомічних даних називають коефіцієнтом "ϕi". Він визначає коефіцієнт кореляції між завданнями з номерами i та j і його обчислюють за формулою

$$\phi_{j,i} = \frac{p_{j,i} - p_j \cdot p_i}{\sqrt{(p_j \cdot q_j) \cdot (p_i \cdot q_i)}}, i = \overline{1,n}; j = \overline{1,m}, \qquad (9)$$

де: $p_{i,j}$ – частка тестованих осіб, що виконали правильно обидва завдання з номерами j та i, тобто частка тих, хто отримав "1" за обома завданнями; p_j – частка тестованих осіб, що правильно виконали одне j-*е* завдання, $q_j = 1 - p_j$, p_i – частка тестованих осіб, що правильно виконали i-*е* завдання тесту, $q_i = 1 - p_i$.

Надійність як внутрішня погодженість визначається зв'язком кожного конкретного елемента вибірки із загальним результатом, тобто тим, наскільки кожен елемент конфліктує з іншими, наскільки кожне окреме запитання тесту вимірює ознаку, на яку скерований весь тест. Для визначення коефіцієнта внутрішньої погодженості розглядають такі методи:

1. Метод розщеплення або метод автономних частин. Можна користуватися однією з таких формул:
   - формула Спірмена-Брауна;
   - формула Рюлона;
   - формула Кьюдера-Річардсона;
   - формула Стенлі.
2. Метод еквівалентних бланків.
3. Метод Альфа-Кронбаха.

Під час застосування методу розщеплення тестову матрицю (результати тестування) розділяють на дві половини, що складаються з парних та непарних номерів завдань.

Для інтерпретації значення коефіцієнта надійності застосовують такі критерії:

Значення коефіцієнта	Інтерпретація
> 0,9	дуже висока надійність
від 0,8 до 0,89	висока надійність тесту
від 0,7 до 0,79	хороша надійність тесту
< 0,7	низька надійність тесту

Використання методу розщеплення дає занижені оцінки надійності через те, що його оцінюють для скороченого у два рази тесту.

У цьому випадку для корекції оцінки надійності, як внутрішньої погодженості відповідно до довжини початкового тесту, можливим є використання формули Спірмена-Брауна, котра має такий вигляд:

$$r_2 = 2 \cdot r_1 / (1 + r_1), \qquad (10)$$

де r_1 – коефіцієнт надійності як стійкості, визначений за формулою Пірсона (7). Для нашого випадку X_i у формулі (7) – це тестовий бал i-го тестованого за запитання з парним номером, Y_i – це тестовий бал i-го тестованого за запитання з непарним номером.

Наведений вище метод оцінки надійності має своєї обмеження у використанні. Його засновано на припущенні паралельності двох половинок тесту, що не завжди і не повною мірою може виявитися правильним. Кореляція двох половинок збільшується із зростанням однорідності тесту.

За результатами проведених розрахунків (табл. 3), значення коефіцієнта надійності як внутрішньої погодженості, вирахуване за формулою Спірмена-Брауна з використанням формули Пірсона, становить 0,76, що свідчить про хорошу надійність тесту. Для підвищення надійності тесту потрібно, за можливості, застосовувати завдання закритого типу, що зменшить вплив суб'єктивізму під час оцінюванні результатів тестування.

Табл. 3. Розрахунок коефіцієнта надійності тестового завдання

Кандидат	Разом правильних відповідей	Експертна оцінка	Надійність (r_1) як стійкість (за формулою Пірсона)	Надійність (r_2) як внутрішня погодженість (за формулою Спірмена-Брауна)	Надійність як внутрішня погоджуваність (за методом Кьюдера-Річардсона (KR-20)
1	5	6	0,6	0,76	0,83
2	13	12			
3	7	8			
4	10	11			
5	8	7			
6	12	12			
7	19	17			
8	9	8			
9	9	8			
…	…	…			
25	5	6			
26	6	7			
27	7	8			
28	6	6			
29	10	12			
Сума	310	304			

Тестові завдання повинні бути коректно сформульовані, інакше сильні студенти можуть пропустити їх, а це негативно вплине на надійність тесту. Надійність тесту зростає відповідно до збільшення кількості запитань у тесті, тому зазвичай вистарчає 30 запитань. Кожен тест потрібно забезпечити чіткою стандартною інструкцією, оскільки її відсутність призведе до неоднозначності, а звідси – і до втрати надійності.

Метод Кьюдера-Річардсона застосовують для оцінки надійності дихотомічних завдань тестів, а також і як метод розщеплення тесту, що заснований на одноразовому тестуванні, але, на відміну від нього, не залежить від штучних допущень про повну паралельність двох частин тіла тесту. Однак сфера його застосування обмежена, оскільки його можна використовувати тільки під час застосування дихотомічних оцінок за результатами виконання гомогенних (тільки за певною дисципліною) тестів.

Формула Кьюдера-Річардсона (KR-20) має такий вигляд:

$$r_{KR-20} = \frac{m}{m-1} \cdot \left(1 - \frac{1}{S_y^2} \sum_{j=1}^{m} p_j q_j\right), \qquad (11)$$

де: p_j – частка правильних відповідей на j-те завдання; $q_j = 1 - p_j$; S_y^2 – дисперсія за сумарним розподілом балів; m – кількість завдань тесту (у нас 30).

Для матриці даних, що представлені в табл. 1, підраховано дисперсію сумарних балів S_y^2=30,1, а долі правильних відповідей отримуємо діленням числа суми правильних відповідей на кількість студентів у групі (у нас 29). Сума добутків часток правильних і неправильних відповідей у такому випадку буде визначатися так: 0,34·0,66+0,21·0,79 + 0,28·0,72 + 0,17·0,83 + + 0,24·0,76 + 0,14·0,86 + 0,21·0,79 + 0,62·0,38 + …+ 0,24·0,76 + 0,62·0,38 + 0,24 + 0,76 = 6,11, а коефіцієнт надійності – $r_{KR\text{-}20}$ =30/29· (1–6,11/30,1)=0,82.

Формула Кьюдера-Річардсона (KR-20) з урахуванням бісеріального коефіцієнта кореляції B_j може мати такий вигляд:

$$p = \frac{s_y^2 - \sum_{j=1}^{m} s_j^2}{2 \cdot s_y^2} + \sqrt{\left(\frac{s_y^2 - \sum_{j=1}^{m} s_j^2}{2 \cdot s_y^2}\right)^2 + \frac{1}{2 \cdot s_y^2} \sum_{j=1}^{m} B_j^2 \cdot s_j^2}, \qquad (12)$$

де: $$B_j = \frac{M_{j,1} - M_{j,0}}{s_y} \cdot \sqrt{\frac{n_{j0} \cdot n_{j1}}{n \cdot (n-1)}}, j = \overline{1,n}; \qquad (13)$$

$n_{j1} = \sum_{i=1}^{n} x_{ij}$ – кількість студентів, що одержали за j-тим завданням один бал; $n_{j0} = n - n_{j1}$ – кількість студентів, що відповіли неправильно на j-те завдання; M_{j1} – середнє арифметичне сум балів по всьому тесту для тих студентів, котрі отримали за цим завданням один бал, M_{j0} – нуль балів:

$$M_{j0} = \frac{1}{n_{j0}} \sum_{i=1}^{n} (1 - x_{i,j}) \cdot y_i, j = \overline{1,m}, \; M_{j1} = \frac{1}{n_{j1}} \sum_{i=1}^{n} x_{i,j} \cdot y_i, j = \overline{1,m}. \qquad (14)$$

Результати розрахунків подано у табл. 4.

Висновки. За результатами проведеного аналізу, завдання, котрі мають індекс дискримінативності D < 0,2 (у нашому випадку це завдання з номерами 5, 28, 29, 30), та ті, що погано корелюють зі сумою балів, тобто B_j < 0,15 (у нашому випадку це завдання 12, 13), повинні бути видалені із збірника тестових завдань. Для зменшеного списку завдань складається нова впорядкована таблиця, для якої перераховуються зазначені вище показники.

Чим вищий показник надійності, тим менша помилка виміру індивідуального результату. Загалом під час оцінювання надійності не можна покладатися тільки на один показник, оскільки кожен із них має свої обме-

ження, що зміщують оцінки надійності тесту в сторону підвищення чи заниження. Для достовірності перевірки якості тесту необхідно враховувати декілька показників надійності, що підраховані за різними формулами (див. табл. 3).

Табл. 4. Розрахунок бісеріального коефіцієнта кореляції тестового завдання

Номер запитання										
Кандидат	1	2	5	…	12	13	29	30	Разом	Експертна оцінка
1	0	0	1	…	0	0	0	0	5	6
2	0	0	0	…	1	1	1	0	12	12
3	0	0	0	…	0	1	1	0	7	8
…	…	…	…	…	…	…	…	…	…	…
25	0	0	0	…	0	1	0	0	5	6
26	0	0	0	…	0	1	1	1	6	7
27	0	0	0	…	0	0	1	0	7	8
28	0	0	0	…	0	1	0	0	6	6
29	0	0	0	…	1	1	0	0	10	12
Всього правильних відповідей	9	6	6	…	13	18	18	7	294	293
Частка правильних відповідей	0,31	0,21	0,21	…	0,45	0,62	0,62	0,24		
Частка неправильних відповідей	0,69	0,79	0,79	…	0,55	0,38	0,38	0,76	середнє значення дисперсії (S_i^2)	дисперсія сумарних балів (S_y^2)
Добуток, дисперсія (S_i^2)	0,21	0,16	0,16	…	0,25	0,24	0,24	0,18	0,19	30,1
M_{i1}	15,2	16,5	12,7	…	11,5	10,5	11,5	15,1	13,83	
M_{i0}	8,32	9,17	10	…	10	11	9,36	9,27	8,83	
Бісеріальний коефіцієнт кореляції (B_i)	0,59	0,55	0,2	…	0,14	-0	0,19	0,47	0,39	

Як нижню межу допустимих значень надійності вибирають зазвичай значення 0,7. За більш низького значення використання тесту не є оправданим через великі похибки вимірювання. Якщо тест розробляють професіонали, то до нього пред'являють жорсткіші вимоги. Зазвичай, тести, що мають надійність меншу 0,8, вважають непридатними у професійно організованих службах і центрах тестування.

Значення коефіцієнта надійності, що перевищують 0,9, свідчать про високу якість тесту. Вони є бажаними, але трапляються дуже рідко. Зазвичай у тестологічній практиці надійність тестів змінюється в інтервалі (0,8:0,9). Маємо надію, що ця робота вдосконалить таку форму перевірки знань, як тестовий контроль із дихотомічними варіантами відповідей, що покращить якість освіти в Україні.

Перелік використаних джерел

Avanesov, V. S. (2005). Teoriia i metodika pedagogicheskikh izmerenii (materialy publikatcii). Moscow: TcT i MKO UGTU-UPI, 98 p. [In Russian].

Bovtrukevich, M. V., & Kireenko, A. V. (2017). Metodika rascheta testovykh kharakteristik. Nauchnoe soobshhestvo studentov: Mezhdistciplinarnye issledovaniia: sb. st. po mater. III mezhdunar. stud. nauch.-prakt. konf., Vol. 3 (pp. 23–28). URL: http://www.sibac.info/sites/default/files/conf/file/stud_3_3.pdf. [In Russian].

Chelyshkova, M. B. (2002). Teoriia i praktika konstruirovaniia pedagogicheskikh testov: uchebnoe posobie. Moscow: Logos, 432 p. Retrieved from: http://www.immsp.kiev.ua/publications/articles/2007/2007_3,4/Fedoruk_034_2007.pdf. [In Russian].

Я. Ф. Кулешнык, В. В. Сеныk, Т. В. Магеровская

Львовский государственный университет внутренних дел, г. Львов, Украина

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ ДЛЯ СОЗДАНИЯ КОГНИТИВНЫХ ТЕСТОВЫХ ЗАДАНИЙ

Подано определение теста, сложности тестового задания, валидности, дискриминативности и надежности тестовых заданий. Обоснована необходимость прохождения тестовыми заданиями процесса специальной оценки с использованием элементов математической статистики. Рассмотрены два основных подхода к созданию тестов и главный вопрос теории тестов – построение оптимального теста. Подана математическая постановка задачи создания и анализа тестовых заданий. На конкретных результатах тестовых заданий показана методика расчёта тестовых характеристик, сложность тестовых заданий рассмотрена на основе эмпирической проверки заданий, с подсчетом части правильных ответов. По известным формулам рассчитаны коэффициенты вариации и дисперсии, сделана оценка валидности с использованием значимого внешнего критерия – экспертной оценки. Подана таблица критериев для интерпретации значения коэффициента валидности. Чтобы сделать невозможным создание некачественных тестовых заданий, для вычисления коэффициента дискриминативности применен метод крайних групп. Приведены критерии интерпретации значения коэффициента дискриминативности. Разработаны конкретные рекомендации для создания тестовых заданий с надлежащим значением коэффициента дискриминативности. Рассмотрены два вида надежности тестов: надежность как стойкость и надежность как внутренняя согласованность. Для определения значения надежности как стойкости применен преобразованный коэффициент корреляции Пирсона "фи" для дихотомических данных. Приведены результаты проведенных расчетов значения коэффициента надежности как внутренней согласованности, вычисленного по формуле Спирмена-Брауна с использованием формулы Пирсона. Подана таблица критериев для интерпретации значения коэффициента надежности. Использован метод Кьюдера-Ричардсона как метод расщепления теста для оценки надежности дихотомических заданий тестов. Дана интерпретация результатов и выработка упрощенных общих рекомендаций составителям, которые можно применять в тестологической практике.

Ключевые слова: адаптивный тест; классические статистические методы; сложность; валидность; дискриминативность; надежность тестовых заданий.

Ya. F. Kuleshnyk, V. V. Senyk, T. V. Maherovska
Lviv State University of Internal Affairs, Lviv, Ukraine

THE USE OF STATISTICAL METHODS FOR CREATION OF COGNITIVE TEST TASKS

In this article, the definitions of test, test task difficulty, validity, discriminatory and test data reliability are given. In addition, the necessity of clarification for test tasks with appropriation of the special valuation, based on mathematical statistics elements is substantiated. Two basic methodologies of test creation are reviewed. Therefore an optimal test creation as the main problem of testing theories is reviewed as well. The mathematical statement of test task creation and test tasks analysis is given; concrete results show the methodology of test characteristics calculations. The difficulty of test tasks is reviewed based on the empirical knowledge checkout with calculation of the part of correct answers. Variation and dispersion coefficients are calculated based on well-known formulas. The valuation of validity is made by using the meaningful external criteria – an expert valuation. Moreover, the criteria table is introduced to interpret the value of validity coefficient. To make the creation of non-correct tests impossible, the method of extreme groups is used to calculate the discriminatory coefficient, and criteria to interpret the discriminatory coefficient data are introduced. Concrete recommendations to create test tasks with appropriate discriminatory coefficient values are developed; two definitions of reliability are reviewed: reliability as stability and reliability as internal consistency. There we used the coefficient of Pearson correlation "ph" to define reliability as stability for dichotomous data, the results of reliability as internal consistency are received using Spearman-Brown formula including Pearson formula, and the criteria table is introduced to interpret the value of reliability coefficient. Richardson-Cuarda Method is used to split test to evaluate reliability of dichotomous test tasks. The interpretation of results is given. A set of simplified common recommendations is provided to compiles, which are possible to use in daily testological practice.

Keywords: adaptive test; classical statistics methods; difficulty; validity; discriminatory test data reliability.