УДК 005.22:37.091.26 **Doi: 10.31767/su. 4(83)2018.04.09**
JEL Classification: C81, I21

В. М. Кухаренко,
кандидат технічних наук,
професор, науковий керівник проблемної лабораторії дистанційного навчання,
Національний технічний університет "Харківський політехнічний інститут",
E-mail: kukharenkovn@gmail.com;
Л. П. Перхун,
кандидат педагогічних наук, доцент,
завідувач сектору дистанційного навчання,
E-mail: lperkhun@gmail.com;
Н. М. Товмаченко,
кандидат технічних наук, старший науковий співробітник,
заступник завідувача сектору дистанційного навчання,
E-mail: avt9tnn@gmail.com;
Національна академія статистики, обліку та аудиту

Методика комплексного оцінювання якості тестів. Частина 2

Продовжено виклад методики комплексного оцінювання якості тестів, що ґрунтується на методах класичної теорії, методах Data Mining та Item Response Theory (IRT). Для аналізу використано коефіцієнт внутрішньої узгодженості Кьюдера – Річардсона, коефіцієнт генералізації, виконана ієрархічна кластеризація, здійснено розрахунки за однопараметричною моделлю Раша. Обґрунтовано засади подальшого вдосконалення тесту.

Ключові слова: *дистанційне навчання, тестове завдання, тест, надійність, двофакторний дисперсійний аналіз, кластеризація, модель Раша.*

У статті [9] було розпочато виклад методики комплексного оцінювання якості тестів, зокрема схарактеризовано перші три кроки з шести виокремлених. Продовжимо виклад, при цьому зазначимо, що формули, таблиці та використані джерела пронумеровано у продовження попередньої статті.

Крок 4. Оцінка надійності тесту

Питання оцінювання надійності тесту виникає при дослідженні якості тесту за класичною теорією тестування. Під надійністю тесту у цій роботі ми розумітимемо ступінь стійкості (повторюваності) і точності результатів вимірювання.

Найбільш поширеними та всебічно висвітленими в літературних джерелах є такі методи оцінювання надійності:

- ретестова надійність – проведення повторного тестування за тими самими тестовими завданнями через певний проміжок часу;
- надійність паралельних форм тесту – проведення повторного тестування за відносно новими, близькими за змістом до початкового варіанта тестовими завданнями;
- надійність частин тесту – розбиття тесту на дві еквівалентні частини, наприклад за парними або непарними номерами тестових завдань, за близькістю значень складності або дискримінативності, за часом виконання тощо [10].

Статистико-математичним підґрунтям оцінювання надійності у перелічених випадках є оцінка міри схожості результатів двох вимірювань – коефіцієнт кореляції Пірсона, коефіцієнт рангової кореляції Спірмена, коефіцієнт рангової кореляції Кендалла тощо.

Недоліком першого підходу є те, що за умов короткого інтервалу часу між сесіями тестування існує велика ймовірність запам'ятовування студентами як окремих питань, так і відповідей на них, що призведе до викривлення оцінок надійності. За умов збільшення інтервалу часу між тестовими випробуваннями відбудеться зміна самої досліджуваної латентної властивості, яку вимірював тест спочатку. Отже, оцінка надійності тесту також не буде достовірною.

Другий підхід за рахунок використання нових тестових завдань дещо мінімізує вплив згадування попередніх відповідей. Однак у цьому випадку на перший план виходить інша проблема. З одного боку, два набори тестових завдань мають бути орієновані на вимірювання або однієї властивості студента (нормативно орієнтоване тестування), або ступеня засвоєння одного й того самого набору знань (критеріально орієнтовані тести). З іншого боку, ці два набори тестових завдань мають бути незалежними.

У третьому підході на перший план виходить питання реальної еквівалентності двох або більше частин тесту. Крім цього, цей підхід ґрунтується на припущенні, що розподіл оцінок тестових завдань у досліджуваному наборі тестів описується нор-

мальним законом. Але це має бути характерним тільки для нормативно орієнованих тестів [11]. Отже, для критеріально орієнтованих тестів цей підхід застосовувати не можна.

На наш погляд, універсальною процедурою оцінювання надійності тестів, яка передбачає одноразове тестування і дозволяє уникнути описаних вище проблем, є така:

1. Оцінювання надійності окремих завдань тесту за коефіцієнтом внутрішньої узгодженості Кьюдера – Річардсона r_{KR-20}:

$$r_{KR-20} = \frac{m}{m-1}\left(1 - \frac{\sum_{j=1}^{m} p_j q_j}{S_x^2}\right), \quad (5)$$

де p_j – частка правильних відповідей на j-те завдання; q_j – частка неправильних відповідей на j-те завдання, $q_j = 1 - p_j$; S_x^2 – дисперсія розподілу спостережуваних індивідуальних балів; $j = \overline{1,m}$, m – кількість тестових завдань.

Значення $0{,}7 \le r_{KR-20} < 0{,}8$ вважають задовільними, $0{,}8 \le r_{KR-20} < 0{,}9$ – добрими, $r_{KR-20} \ge 0{,}9$ – відмінними [12]. Слід зауважити, що цей коефіцієнт застосовують тільки у випадку дихотомічних (бінарних) оцінок тестових завдань.

2. Оцінювання надійності тесту загалом (усього набору оцінок) за коефіцієнтом генералізації [10; 13]:

$$\rho_{\bar{Y}}^2 = \frac{s_1^2}{s_1^2 + (s_2^2 + s_{error}^2)/m}, \quad (6)$$

де S_1^2 – дисперсія оцінок студентів; S_2^2 – дисперсія складності завдання; s_{error}^2 – дисперсія похибки вимірювання за усіма тестовими завданнями.

Як правило, тести з надійністю, меншою за 0,8, вважаються непридатними у професійно організованих службах і центрах тестування. Значення коефіцієнта надійності, що перевищують 0,9, говорять про високу якість тесту. Зазвичай в тестологічній практиці надійність тестів приймається як задовільна на рівні 0,8 – 0,9.

Для розрахунку коефіцієнта генералізації можуть бути використані результати двофакторного дисперсійного аналізу ANOVA без повторних вимірювань. У термінах ANOVA-аналізу факторами будуть рівні підготовки студентів та рівні складності тестових питань.

Розрахуємо коефіцієнт надійності Кьюдера – Річардсона за формулою (5):

$$r_{KR-20} = \frac{10}{10-1}\left(1 - \frac{1{,}9}{6{,}89}\right) = 0{,}79 \approx 0{,}8. \quad (7)$$

Обчислений нами коефіцієнт свідчить про задовільну надійність окремих завдань тесту.

Для розрахунків за коефіцієнтом генералізації наведемо результати двофакторного дисперсійного аналізу ANOVA без повторних вимірювань, виконаного у пакеті статистичних програм (ПСП) SPSS (табл. 7, за даними матриці (1)).

Таблиця 7

Результати двофакторного дисперсійного аналізу без повторних вимірювань

Джерело варіації	Сума квадратів, *SS*	Ступені свободи, *df*	Середній квадрат, *MS*	Критерій Фішера		*p*-значення
				F-розрахункове	*F*-критичне	
Тестовані, s_1^2	6,200	$n-1 = 9$	$MS_1 = 0{,}689$	4,359	3,79	0,000
Тестові завдання, s_2^2	6,000	$m-1 = 9$	$MS_2 = 0{,}667$	4,219	3,79	0,000
Похибка, s_{error}^2	12,800	$n(m-1) = 81$	$MS_{error} = 0{,}158$			
Разом	25,000	$N-1=99$				

Значення рівня значущості *F*-критерію $p<0{,}05$ у табл. 7 дозволяє прийняти гіпотезу про статистично значущий вплив на індивідуальний бал студента (тобто на результати виконання тесту загалом) фактора рівня підготовки студентів та фактора складності тестових завдань.

Наведемо результати обчислень за формулою (6) з використанням даних табл. 7:

$$s_1^2 = (MS_1 - MS_{error}) = (0{,}689 - 0{,}158) = 0{,}531; \quad (8)$$

$$s_2^2 = (MS_2 - MS_{error}) = (0{,}667 - 0{,}158) = 0{,}509; \quad (9)$$

$$s_{error}^2 = MS_{error}/(n(m-1)) = 0{,}158; \quad (10)$$

$$\rho_{\bar{Y}}^2 = \frac{0{,}531}{(0{,}531 + (0{,}509 + 0{,}158)/10)} = 0{,}888 \quad (11)$$

Розраховане значення коефіцієнта генералізації для тесту за даними матриці (1) дорівнює 0,888, що свідчить про високу надійність тесту в цілому.

Крок 5. Диференціація студентів

Одним із завдань застосовування тестового контролю є диференціація студентів за їх рівнем засвоєння навчального матеріалу. Подібний аналіз рекомендується здійснювати ієрархічними кластерними методами, наприклад методом Уорда, при якому всередині кластерів оптимізується мінімальна дисперсія і в результаті формуються кластери приблизно рівних розмірів. За міру відмінності між кластерами обирається квадратична евклідова відстань, що сприяє збільшенню їх контрастності. Цей метод реалізовано авторами за допомогою пакета SPSS, отримані результати подано на рис. 4 (побудовано за даними матриці (1)).

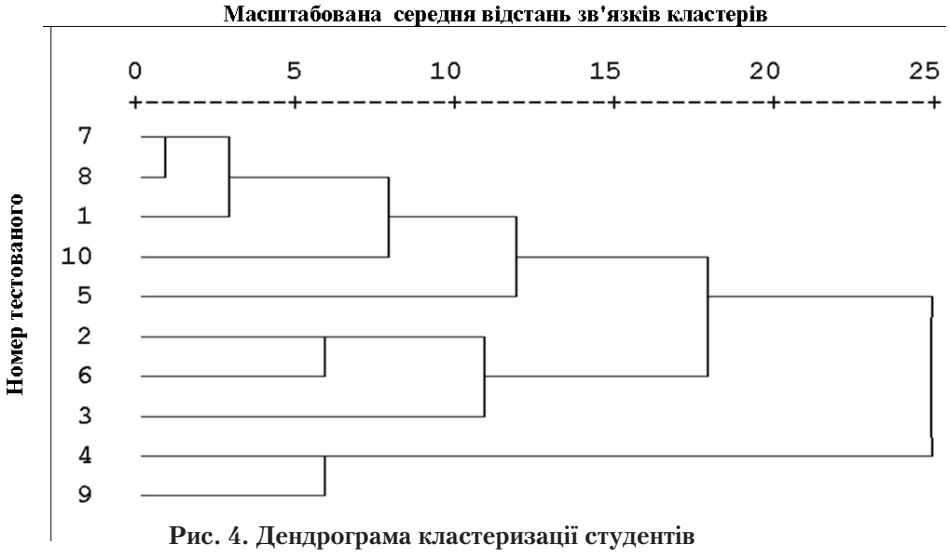


Рис. 4. Дендрограма кластеризації студентів

За результатами тестування студентів можна розбити на три групи. До першої віднесено тестованих з номерами 7, 8, 1, 10, 5; до другої – 2, 6, 3; до третьої – 4, 9. У табл. 8 подана розширена інформація щодо членів кожного кластеру із зазначенням індивідуального бала кожного.

Таблиця 8

Розподіл студентів за кластерами

Тестовані	Індивідуальний бал тестованого	Належність тестованих до кластера ієрархічного кластерного аналізу
1	6,0	2
2	2,0	1
3	1,0	1
4	9,0	3
5	4,0	2
6	4,0	1
7	5,0	2
8	4,0	2
9	9,0	3
10	6,0	2

У роботі [14] авторами описаний алгоритм виявлення тестових завдань, які мають здатність диференціювати студентів за рівнем засвоєння теми / дистанційного курсу, а також реалізація цього алгоритму на основі дерев класифікацій засобами ПСП Statistica. Там же проілюстрована побудова класифікаційних дискримінантних функції засобами ПСП SPSS, які дозволяють віднести студента до групи з певним рівнем знань (іншими словами – до певного кластера).

Крок 6. Конструювання оптимальних тестів засобами Item Response Theory

Основним базовим постулатом для моделей Item Response Theory (IRT) є твердження про те, що якісний тест складається тільки з тих тестових завдань, які відповідають конкретній моделі вимірювання. Тобто тест має "вписатись у модель" [14]. У сучасній теорії тестування припускається, що результат виконання тесту залежить від двох латентних характеристик: рівня трудності тестового завдання (b) та рівня засвоєння навчального матеріалу студентом (θ). Обидві величини вимірюються у логітах. Логіт b розраховується як натуральний логарифм від ділення часток не-

правильних і правильних відповідей студентів на певне тестове запитання. Логіт θ розраховується як натуральний логарифм від ділення часток правильних і неправильних відповідей студента на всі завдання тесту.

На сьогодні у межах сучасної теорії тестування розроблено кілька моделей тестування, зокрема двопараметрична та трипараметрична моделі Бірнбаума, які включають, окрім згаданих раніше, додаткові параметри: складність тестового завдання, дискримінативність тестового завдання, угадування студентом правильної відповіді [15; 16]. Водночас класична однопараметрична модель Раша показує аналогічні з названими моделями результати, але потребує меншої кількості вхідних параметрів і розрахунків.

Класична логістична однопараметрична модель Раша задається формулами:

$$P_j(\theta) = \frac{\exp(1{,}7\ (\theta - b_j))}{1 + \exp(1{,}7\ (\theta - b_j))}, \qquad (12)$$

$$P_i(b) = \frac{\exp(1{,}7\ (\theta_i - b))}{1 + \exp(1{,}7\ (\theta_i - b))}. \qquad (13)$$

де $P_j(\theta)$ – ймовірність правильної відповіді студентами на *j*-те тестове завдання; $P_i(b)$ – ймовірність успішного виконання *i*-м студенту тесту загалом; θ_i – рівень підготовленості того, хто тестується; b_j – рівень складності *j*-го тестового завдання.

Оскільки ймовірність успішної відповіді залежить, по суті, тільки від різниці $\theta - b$, модель вважається однопараметричною.

Функція $P_j(\theta)$ є зростаючою функцією правильності виконання тестового завдання: чим вище рівень підготовленості студента, тим більша ймовірність правильного виконання ним *j*-го завдання тесту. Графік цієї функції називається характеристичною індивідуальною кривою *j*-го тестового завдання. Функція $P_i(b)$ є спадною функцією успішного виконання і-м студентом тесту загалом: чим складніше будуть завдання, тим менша ймовірність успішного виконання студентом тесту загалом. Графік цієї функції називається характеристичною індивідуальною кривою *i*-го студента.

Для розрахунку вхідних даних для моделі Раша користуються такими формулами.

$$p_i = \frac{\sum_{j=1}^{m} x_{ij}}{m}, i = \overline{1,n}, \qquad (14)$$

де p_i – частка правильних відповідей *i*-го студента на всі завдання тесту; x_{ij}– варіанти відповідей *i*-го студента на *j*-те завдання тесту (за даними матриці (1)); *n* – кількість студентів;

$$q_i = 1 - p_i\ , \qquad (15)$$

де q_i – частка неправильних відповідей і-го студента на всі завдання тесту;

$$\theta_i^0 = \ln(p_i / q_i), i = \overline{1,n}\ , \qquad (16)$$

де θ_i^0 – початковий рівень підготовленості студента;

$$b_j^0 = \ln(q_j / p_j), j = \overline{1,m}\ , \qquad (17)$$

де b_j^0 – початковий рівень складності тестового завдання.

Отримані значення θ_i і b_j дозволяють зіставити рівень знань студентів із рівнем складності завдань тесту. Якщо $\theta_i - b_j$ – від'ємна величина, то завдання складності b_j є надто важким для студента з рівнем знань θ_i. Якщо ця різниця додатна, то завдання надто легке.

Для запису оцінок узагальнених параметрів θ і b в єдиній інтервальній шкалі необхідно їх скоригувати на відповідні дисперсії. Детальний опис розрахунків можна знайти в роботі [17]. Слід зауважити, що сума параметрів b, зведених у єдину шкалу, має прямувати до нуля. Занадто велике додатне значення свідчить про високу складність тесту або про низьку підготовку групи студентів. Від'ємне значення тлумачиться навпаки. Результати обчислень $P_j(\theta)$ та $P_i(b)$ наведено у табл. 9 (розраховані за даними матриці (1)).

Експериментальні характеристичні криві для тестових завдань наведено на рис. 5, для окремих студентів – на рис. 6 (обидва рисунки побудовано у середовищі MathCad за даними матриці (1)). На рис. 5 характеристичні криві впорядковані за зростанням аргументу *b*, тобто завдання тесту в порядку зростання складності. Характеристичні криві для тестових завдань 5 і 6 збіглися. Характеристична крива *j*-го завдання тесту показує взаємозв'язок між значеннями незалежної змінної θ і значеннями p_j. Точці перегину характеристичної кривої відповідають координати $\theta = b$, а p_j у цій точці дорівнює 0,5.

Як видно з рис. 5, тестові питання "вписались" у модель Раша. Графіки логістичних кривих розташовуються від найлегшого до найскладнішого зліва направо. Однак цей тест не є якісним згідно з теорією IRT, оскільки:

– деякі характеристичні криві наклалися одна на одну. Це означає, що відповідні тестові питання мають однакову складність, вони мають бути видалені (нормативно орієнтований тест) або реконструйовані (критеріально орієнтований тест). Визначити, на якому саме тестовому питанні слід акцентувати увагу, можна за допомогою двопараметричної моделі Бірнбаума. Більш детальне обґрунтування доцільності модифікації або ви-

Таблиця 9

Значення $P_j(\theta)$ та $P_i(b)$

Вхідні дані для розрахунку		b_j^0									
		Номер завдання									
Номер тестованого	θ_i^0	1	2	3	4	5	6	7	8	9	10
		–2,197	**–1,386**	**–0,847**	**–0,405**	**0**	**0**	**0,4057**	**0,847**	**1,386**	**2,197**
1	0,405	0,988	0,955	0,894	0,799	0,666	0,666	0,500	0,321	0,159	0,045
2	–1,386	0,799	0,500	0,286	0,159	0,087	0,087	0,045	0,022	0,009	0,002
3	–2,197	0,500	0,201	0,092	0,045	0,023	0,023	0,012	0,006	0,002	0,001
4	2,197	0,999	0,998	0,994	0,988	0,977	0,977	0,955	0,908	0,799	0,500
5	–0,405	0,955	0,841	0,679	0,500	0,334	0,334	0,201	0,106	0,045	0,012
6	–0,405	0,955	0,841	0,679	0,500	0,334	0,334	0,201	0,106	0,045	0,012
7	0,000	0,977	0,913	0,809	0,666	0,500	0,500	0,334	0,191	0,087	0,023
8	–0,405	0,955	0,841	0,679	0,500	0,334	0,334	0,201	0,106	0,045	0,012
9	2,197	0,999	0,998	0,994	0,988	0,977	0,977	0,955	0,908	0,799	0,500
10	0,405	0,988	0,955	0,894	0,799	0,666	0,666	0,500	0,321	0,159	0,045

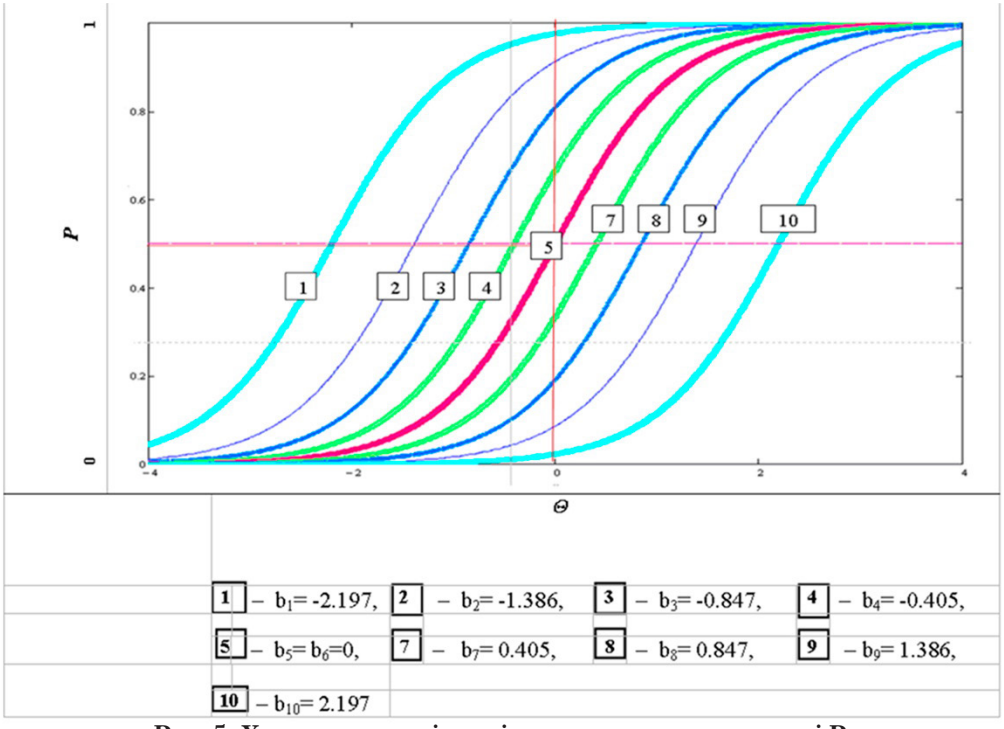
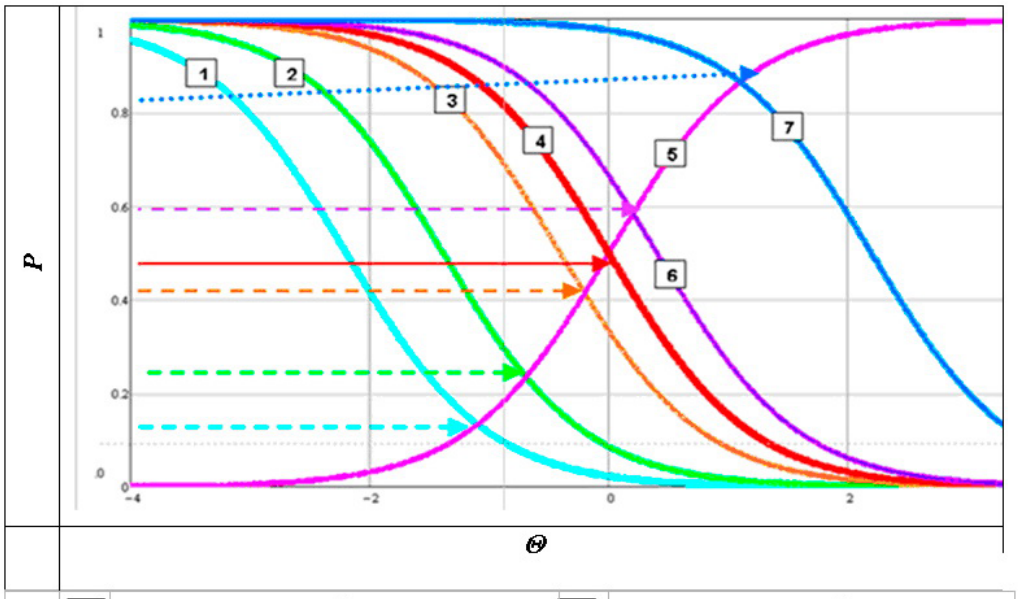


Рис. 5. Характеристичні криві тестових завдань у моделі Раша

лучення окремого тестового завдання наведено у роботі [18];

– щільність характеристичних кривих неоднакова; у тест необхідно додати завдання або змінити тестові завдання-дублікати у такий спосіб, щоб заповнити прогалини осі абсцис, де немає характеристичних кривих.

Характеристичні криві студентів (див. рис. 6), побудовані за моделлю Раша, дозволяють поділити останніх на дві групи – сильні (мають додатні логіти) і слабкі (мають від'ємні логіти). На рисунку чітко видно, що чим вище рівень підготовленості студента, тим більша ймовірність правильної відповіді на тестове питання 5 (точка перетину логістичних кривих 5 і 7 має найбільшу ординату з усіх точок перетину). Аналогічно можна визначити ймовірність правильної відповіді студента на кожне запитання.



	1	– тестований № 3, Θ= -2.197,	2	– тестований № 2, Θ= -1.386,
	3	– тестовані № 5, № 6, № 8, Θ= -0.405,	4	– тестований № 7, Θ= 0,
	5	– тестові завдання 5, 6, Θ= 0,	6	– тестовані № 1, № 10, Θ= 0.405,
	7	– тестовані № 4, № 9, Θ= 2.197		

Рис. 6. Характеристичні криві студентів у моделі Раша

Запропонована авторами методика комплексного оцінювання якості тестів, що ґрунтується на методах класичної теорії, методах Data Mining та Item Response Theory, дозволяє вивчати блок тестових завдань з різних позицій. Зокрема, оцінювати окремі тестові завдання, тест загалом, окремого студента, групу загалом тощо. У контексті практичного впровадження цієї методики корисно розробити окремий плагін, сумісний із платформою дистанційного навчання Moodle, який би автоматизував описані розрахунки. Адже підґрунтя для цього вже є – певна статистика щодо виконаних тестових завдань збирається і навіть обраховується [3; 9]. У теоретичному плані автори зацікавлені у вивченні меж застосування двопараметричної та трипараметричної моделей Бірнбаума для удосконалення процесу та результату тестування студентів у системах дистанційного навчання.

Список використаних джерел

9. Кухаренко В. М., Перхун Л. П., Товмаченко Н. М. Методика комплексного оцінювання якості тестів. Частина 1 // Статистика України. 2018. № 3. С. 40–48.

10. Морозов С. М. Засоби контролю діагностичних якостей психологічних тестів: навч. посібник. Київ: ІСДО, 1994. 68 с. URL: https://psyfactor.org/lib/morozov1.htm

11. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов. Москва: Логос, 2010. 668 с.

12. Сіницький М. Є. Статистичні інструменти вимірювання якості освіти. Частина 2. Класичний підхід // Науковий вісник НАСОА. 2015. № 1. С. 75–86.

13. Ковальчук Ю. О. Теорія освітніх вимірювань. Ніжин: Вид. ПП Лисенко М. М., 2012. 200 с.

14. Кухаренко В., Перхун Л., Товмаченко Н. Тестовий контроль знань: інструменти інтелектуального аналізу та Item Response Theory // Інноваційні комп'ютерні технології у вищій школі: мат. 10-ї наук.-практ. конф. (21–23 листопада 2018 р., м. Львів) / відп. за вип. Л. Д. Озірковський. Львів: Видавництво Львівської політехніки, 2018. С. 71–76.

15. Федорук П. І.Адаптивні тести: статистичні методи аналізу результатів тестового контролю знань // Математичні машини і системи. 2007. № 3. С. 122–138. URL: http://www.immsp.kiev.ua/publications/articles/2007/2007_3,4/Fedoruk_034_2007.pdf

16. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: The University of Chicago Press., 1980. 199 p.

17. Ким В. С. Тестирование учебных достижений: монография. Уссурийск: Изд-во УГПИ, 2007. 214 с. URL: http://clipperkim.narod.ru/test/monotest/index.html

18. Челышкова М. Б. Теория и практика конструирования педагогических тестов: учеб. пособие. Москва: Логос, 2002. 432 с.

References

9. Kukharenko, V. M., Perkhun, L P., & Tovmachenko, N. M. (2018). Metodyka kompleksnoho otsiniuvannia testiv. Ch. 1 [The Method for Comprehensive Quality Evaluation of Tests. Part 1]. *Statystyka Ukrainy – Statistics of Ukraine, 3,* 40–48 [in Ukrainian].

10. Morozov, S. M. (1994). *Zasoby kontrolju diagnostychnych jakostej psychologichnych testiv [Means of control of diagnostic properties of psychological tests].* Kyiv: ISDO. Retrieved from https://psyfactor.org/lib/morozov3.htm [in Ukrainian].

11. Kroker, L., & Algina J. (2010). *Vvedenie v klassicheskuiu i sovremennuiu teoriiu testov* [Introduction to Classical and Modern Test Theory]. Moscow: Logos [in Russian].

12. Sinytskyi, M. Ye. (2015). Statystychni insrumenty vymiriuvannia yakosti osvity. Ch. 2. Klassychnyi pidchid [Statistical tools for measuring the quality of education. Part 2. Classical approuch]. *Naukovyi visnyk Natsionalnoi akademii statystyky, obliku ta audytu – Scientific Bulletin of the National Academy of Statistics, Accounting and Audit, 1,* 75–86 [in Ukrainian].

13. Kovalchuk, Yu. O. (2012) *Teoriia osvitnich vumiriuvan [Theory of educational measurements].* Nizhyn: Vydavets PP Lysenko M. M. [in Ukrainian].

14. Kukharenko, V. M., Perkhun, L P., & Tovmachenko, N. M. (2018). Testovyi kontrol znan: instrumenty intelektualnoho analizu ta Item Response Theory [Test Knowledge Control: Tools for Intellectual Analysis and Item Response Theory]. Proceedings from Innovative Computer Technologies in Higher School: *Desiata naukovo-praktychna konferentsiia (21–23 lystopada 2018 hoda) – Tenth Scientific and Practical Conference.* (pp. 71–78). Lviv: Lviv Polytechnic Publishing [in Ukrainian].

15. Fedorchuk, P. I. (2007). Adaptyvni testy: statystychni metody analizu rezultativ testovogo kontroliu znan [Adaptive tests: statistical methods for analyzing results of the test knowledge control]. *Matematychni mashyny i systemy – Mathematical Machines and Systems, 3*, 122–138. Retrieved from http://www.immsp.kiev.ua/publications/articles/2007/2007_3,4/Fedoruk_034_2007.pdf [in Ukrainian].

16. Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: The University of Chicago Press [in English].

17. Kim, V. S. (2007). *Testirovanie uchebnych dostizhenii [Learning Achievement Testing].* Ussuriisk: UGPI. Retrieved from http://clipperkim.narod.ru/test/monotest/index.html [in Russian].

18. Shelyschkova, M. B. (2002). *Teoriia i praktika konstruirovaniia pedahgohgicheskich testov [Theory and practice of constructing pedagogical tests* . Moscow]: Logos [in Russian].

В. Н. Кухаренко,
кандидат технических наук,
профессор, научный руководитель
проблемной лаборатории дистанционного обучения,
Национальный технический университет
“Харьковский политехнический институт”;
Л. П. Перхун,
кандидат педагогических наук, доцент,
заведующая сектором дистанционного обучения;
Н. Н. Товмаченко,
кандидат технических наук, старший научный сотрудник,
заместитель заведующей сектором дистанционного обучения;
Национальная академия статистики, учета и аудита

Методика комплексного оценивания качества тестов. Часть 2

Продолжено изложение методики комплексного оценивания качества тестов, которая базируется на методах классической теории, методах Data Mining и Item Response Theory (IRT). При анализе

использованы коэффициент внутренней согласованности Кьюдера-Ричардсона, коэффициент генерализации, иерархическая кластеризация, выполнены расчеты по однопараметрической модели Раша. Обоснованы принципы дальнейшего совершенствования теста.

Ключевые слова: *дистанционное обучение, тестовое задание, тест, надежность, двухфакторный дисперсионный анализ, кластеризация, модель Раша.*

V. M. Kukharenko,
PhD in Engineering,
Professor, Scientific supervisor of Problem laboratory of distance learning,
National Technical University "Kharkiv Polytechnic Institute";
L. P. Perkhun,
PhD in Pedagogy, Associate Professor,
Head of the Distance Learning Sector;
N. M. Tovmachenko,
PhD in Engineering, Senior Researcher,
Deputy Head of the Distance Learning Sector;
National Academy of Statistics, Accounting and Audit

The Method for Comprehensive Quality Evaluation of Tests. Part 2

In the article, the description of the complex evaluation method is given, as well as the classical method of Data Mining and Item Response Theory (IRT). In the general method there are six steps. This article describes steps 4-6.

The fourth step of the method is to evaluate the reliability of the test. A universal two-step procedure is proposed – the assessment of the reliability of individual test tasks based on the coefficient of internal coherence of Kjuder – Richardson and the evaluation of the reliability of the test as a whole by the coefficient of generalization. The first of the coefficients is considered acceptable at the level of 0.7 and above, the second – at the level of 0.8 and above. Two-factor ANOVA variance analysis without repeated measurements in SPSS was used to calculate the second coefficient.

At the fifth stage of the methodology, the quality of students' differentiation is assessed by a test that is being studied. The tool for this is selected hierarchical cluster procedures, classification trees and classification discriminant functions. The calculations were performed by means of Statistica and SPSS. Three clusters of students with high, medium and low academic performance were identified. It is shown that the test under study allows the differentiation of students.

At the last, sixth stage, a study of the quality of the test is described based on the one-parameter model of Rash. The levels of the difficulty of the test assignment and the mastering of the student's study material are measured in logics. The analytical task of the characteristic individual curve of the test assignment and the characteristic individual curve of the student, as well as the auxiliary formulas for their calculations, are given. The description is illustrated by a specific example. It is noted that the characteristic curves of students based on the Rash model by means of MathCAD, can clearly divide the latter into two groups – strong (have positive logic) and weak (have negative logic). Recommendations on the interpretation of the obtained results for certain test tasks are formulated. In particular, in case of overlap of the characteristic curves of various test tasks, they must be deleted (normative-oriented test) or reconstructed (criterion-oriented test). This paper does not consider how to determine which test question is to be deleted or corrected, but it is indicated that this can be established with the help of a two-parameter Birnbaum model. If the density of the characteristic curves of the test tasks is not the same; It is recommended to add a test task (in the case of a normative-oriented test) or thus change the duplicate test questions (in the case of a normative-oriented test) to fill the gaps of the abscissa, where there are no characteristic curves.

By the practical implementation of this technique, the authors determine the development of a separate plug-in that is compatible with the Moodle distance learning platform.

The prospect of further research in the theoretical framework is determined by the authors of the study of the boundaries of the use of two-parameter and three-parameter models of Birnbaum to improve the process and test results of students in distance learning systems.

Key words: *distance learning, test task, test, reliability, two-factor variance analysis, clusterization, Rash model.*

Бібліографічний опис для цитування:

Кухаренко В. М., Перхун Л. П., Товмаченко Н. М. Методика комплексного оцінювання якості тестів. Частина 2 // Статистика України. 2018. № 4. С. 72–79.