педагогічними вузами і школами. Співпраця викладачів кафедри та шкільних вчителів має бути спрямована в основному на забезпечення: системності, яка розкривається в цілісному поєднанні мети, завдань, форм і методів співпраці між вчителями школи та викладачами педагогічного університету; систематичності і послідовності визначених напрямів діяльності; поєднання теоретичної та практичної діяльності. Успіх спільної роботи також залежить від вибору шкіл-партнерів для організації взаємодії з майбутніми, початківцями та досвідченими педагогами-професіоналами.

Відповідна організація навчання може бути ефективним засобом підвищення конкурентоспроможності майбутнього вчителя, сприятиме глибокому і різнобічному професійному розвитку як майбутніх вчителів математики так учителів практиків.

Спиок використаних джерел:

1) Матяш О.І. Педагогічна практика в школі: Посібн. для студ. IV-V к. фізико-математ. Факультету/ О.І.Матяш. – Вінниця: ВДПУ, 1999.- 50 с.

2) Михайленко Л.Ф. До питання організації індивідуальної роботи студентів/ Л.Ф.Михайленко // Дидактика математики: проблеми і дослідження. Міжнародний збірник наукових робіт. – Випуск 28. – Донецьк: ДонНУ, 2007.-С.34-36.

3) Михайленко Л.Ф. До питання організації та оцінювання пропедевтичної педагогічної практики студентів математичних спеціальностей в умовах кредитно-модульного навчання / Л.Ф.Михайленко, О.І. Матяш // Сучасні інформаційні технології та інноваційні методики навчання в підготовці фахівців: методологія, теорія, досвід, проблеми. Збірник наукових праць Випуск . /Редкол.: І.А. Зязюн (голова) та ін. – Київ – Вінниця: ДОВ Вінниця, 2010.- С.

УДК 519.8:330.332

Д. А. Найко, Вінниця, Україна / Dm. Naiko, Vinnytsia, Ukraine
e-mail: dmnaiko@ukr.net

ОЦІНКА ЯКОСТІ ЗАКРИТИХ МАТЕМАТИЧНИХ ТЕСТІВ

***Анотація**. У статті розглядаються питання якості тестових завдань та якості дистракторів у закритих математичних тестах. З різних точок зору аналізуються числові характеристики,через які проводиться оцінювання тестів та тестових завдань. Висвітлюється питання доцільності розподілу завдань тесту (за складністю) за нормальним законом. Розкривається логіка застосування того чи іншого підходу при конструюванні тестів чи аналізі самих результатів тестування. Поняття кореляційного зв'язку розглядається на прикладах вивчення зв'язків між різними показниками освітніх вимірювань. Широко обговорюється поняття валідності, яке є одним з найфундаментальніших понять тестології. У статті висвітлюються різні трактування цього поняття. Розглядається валідність методу та валідність тесту. В роботі також розкривається суть прогностичної валідності як виразника зв'язку результатів тестування з прогнозними досягненнями учасників тестування. Така характеристика тесту як надійність є показником точності освітніх вимірювань і аналізується у порівнянні з поняттям валідності тесту. Обговорюється метод порогових груп, найважливіша тенденція якого полягає у тому, що найсильніші групи екзаменованих осіб найчастіше вибирають правильну відповідь, а всі дистрактори, якщо вони якісні, мають приблизно однакову кількість «голосів».*

***Ключові слова**. Тест, тестові завдання, дистрактори, складність тесту, коефіцієнт кореляції, валідність тесту, надійність тесту.*

***Abstract.** The article deals with questions of the quality of test tasks and the quality of distractors in closed mathematical tests. From different perspectives, numerical characteristics,through which tests and test assignments are evaluated, are analyzed. The question of the expediency of the assignment of the test tasks (by complexity) is covered by the normal law. The logic of the application of one or another approach is revealed in constructing tests or analyzing the results of testing itself. The concept of correlation communication is considered on examples of studying the links between different indicators of educational measurements. The concept of validity, which is one of the most*

fundamental concepts of theology,is widely discussed. The article covers various interpretations of this concept. The validity of the method and validity of the test are considered. The paper also reveals the essence of the prognostic validity as the link between the results of testing and the predictive achievements of test participants. Such a test characteristic as reliability is an indicator of the accuracy of educational measurements and is analyzed in comparison with the concept of the validity of the test. The method of threshold groups is discussed, the most important trend of which is that the strongest groups of examiners most often choose the correct answer, and all distractors, if they are qualitative, have roughly the same number of "votes".

Keywords. *Test, test tasks, distractors, test complexity, correlation coefficient, test validity, test reliability.*

Вступ. У роботі [1] вказано на одну з можливостей ширшого впровадження тестування при оцінюванні знань з математичних дисциплін. Цей підхід пов'язаний з побудовою так званих каркасних варіантів відповідей. Тепер ми певною мірою зосереджуємо свою увагу на методах оцінювання якості математичних тестів. Розглядаються питання якості тестових завдань та характеру розподілу дистракторів.

Головна частина. Ми розглядаємо закриті математичні тести з однією правильною відповіддю, решта – дистрактори.

1. *Точкові оцінки якості тестових завдань та тесту.* Зазвичай нарахування балів за окремі тестові завдання проводиться за дихотомічною схемою: за правильну відповідь на завдання виставляється оцінка 1, за неправильну – оцінка 0.

Нами було проведено тестування 25 осіб. Тест складався з 10 завдань, до кожного з яких пропонувалося 4 варіанти відповіді, одна з них − правильна. Результати тестування подано у вигляді таблиці 1.

Таблиця 1

Номер учасника	Номер тестового завдання										Σ
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	1	1	0	1	0	1	8
2	0	0	0	1	1	0	0	1	1	1	5
3	0	0	0	0	1	0	0	0	0	0	1
4	0	1	0	0	1	0	0	0	0	0	2
5	0	0	0	0	1	0	0	0	0	0	1
6	0	0	0	0	1	0	0	1	1	0	3
7	0	0	0	0	0	0	0	0	0	0	0
8	1	1	1	1	1	1	1	0	1	1	9
9	1	1	1	0	1	1	0	1	1	1	8
10	1	0	0	0	1	0	0	0	0	1	3
11	0	1	0	0	1	0	0	0	0	0	2
12	1	1	0	1	1	0	0	0	1	0	5
13	0	0	0	0	1	0	0	0	0	0	1
14	0	1	0	1	0	1	0	1	1	0	5
15	1	0	0	1	1	0	0	1	0	0	4
16	0	1	0	0	1	0	0	1	1	0	4
17	1	1	0	1	1	0	0	1	1	1	7
18	0	0	0	0	1	0	0	1	1	0	3
19	1	0	0	1	1	0	0	0	0	0	3
20	1	1	1	1	1	1	1	1	1	1	10
21	1	1	0	0	1	1	0	0	0	0	4
22	0	1	0	1	1	0	0	0	0	1	4
23	0	0	0	1	1	0	0	0	0	0	2
24	1	1	0	1	1	0	0	0	1	1	6
25	1	1	0	1	1	1	0	1	1	0	7
Σ	12	14	4	13	23	7	2	11	12	9	107

Таблиця 1 є дихотомічною матрицею A розмірності 25×10, елементи a_{ij} якої дорівнюють 1 або 0.

Якщо випадкова величина X – кількість балів, набрана будь-яким з учасників тестування, то розподіл частот n_i значень цієї величини має вигляд:

X	0	1	2	3	4	5	6	7	8	9	10
n_i	1	3	3	4	4	2	1	2	2	1	1

Розподіл відносних частот задається рядом:

X	0	1	2	3	4	5	6	7	8	9	10
w_i	1/25	3/25	3/25	425	4/25	2/25	1/25	2/25	2/25	1/25	1/25

Часто для візуального порівняння ряд частот та ряд відносних частот доцільно подавати графічно у вигляді *полігона* чи *гістограми* частот.

Основними точковими (числовими) оцінками ряду розподілу частот є *вибіркове середнє* (математичне сподівання) $\overline{X}$, *незміщена вибіркова дисперсія* S_X^2, *стандартне відхилення* (середнє квадратичне відхилення) S_X, *мода*, *медіана*, *коефіцієнт асиметрії* та *коефіцієнт ексцесу*.

Ці числові характеристики дають змогу оцінювати тест з певної точки зору, проте до висновків щодо якості тесту треба підходити дуже обережно. Адже, наприклад, причиною великої дисперсії (особливо для малих вибірок) може бути не лише неякісне складання завдань тесту, але й велика неоднорідність групи осіб, що підлягають тестуванню [2]. Тому велика дисперсія не повинна спонукати розробника тесту до негайної заміни тестових завдань [3, с.131]. Тест з навчальної дисципліни має орієнтуватися на стандарт, яким є програма дисципліни, а не «причісуватися» до конкретного рівня тестувальників. У цьому контексті важливого значення набуває така характеристика, як валідність тесту, про що мова ітиме далі.

Нехай дихотомічна матриця $A=(a_{ij})$ має розмірність $N\times M$. Тоді число $X_i=\sum_{j=1}^{M} a_{ij}$ є *індивідуальним балом* i-го учасника тестування. Число отриманих правильних відповідей на j-е завдання тесту дорівнює $R_j=\sum_{i=1}^{N} a_{ij}$; $N-R_j$ – число отриманих неправильних відповідей на j-е завдання. Таким чином, частка правильних відповідей на j-е завдання дорівнює $p_j=\frac{R_j}{N}$, а число $q_j=1-p_j$ – частка неправильних відповідей на j-е завдання ($0\le p_j\le 1,\quad 0\le q_j\le 1$). У класичній теорії тестування число p_j називають *складністю* j-го завдання. Зрозуміло, що коли p_j близьке до 1, то j-е завдання виявилося нескладним для учасників тестування, якщо p_j близьке до 0, то j-е завдання виявилося складним.

У таблиці 1 варто звернути увагу на такі два моменти: із завданням № 5 успішно впоралися 92% учасників тестування, а із завданням № 7 – лише 8% учасників. Таку ситуацію треба аналізувати, щоб обґрунтовано вилучати або не вилучати з тесту завдання такого типу, бо якісно побудований тест повинен містити зовсім мало нескладних завдань та зовсім мало складних.

На практиці рівень складності як найпростіших так і найскладніших завдань визначається розробником тестів, в залежності від рівня загальної підготовленості групи осіб, що підлягають тестуванню. В усякому разі за складністю завдання тесту повинні розподілятися за нормальним законом: складність більшості з них повинна бути ближче до числа 0,5, аніж до 0 чи до 1.

Завдання тесту, які успішно проходять усі учасники тестування або які не проходить ніхто, зазвичай вилучаються з тесту, оскільки такі завдання не дають можливості диференціювати учасників.

Дисперсія σ^2 як показник розсіювання значень випадкової величини, що набуває значень 0 або 1, визначається за формулою $\sigma^2=p_jq_j$. Оскільки найбільша диференціація учасників тестування на слабших та сильніших здійснюється через завдання середньої складності (для них $p_j=0,5,\ \ q_j=0,5$), то звідси випливає, що «найбажаніша» дисперсія для більшості тестових завдань має дорівнювати 0,25.

Якщо повернутися до таблиці 1, то дисперсії відповідних тестових завдань наведено у наступній таблиці 2.

Бачимо, що з точки зору дисперсійного аналізу наш тест є досить якісним.

Крім того, середня складність завдань тесту дорівнює

$\frac{1}{10}\sum p_j = 0,428 \approx 0,5$, що також підтверджує збалансованість тесту.

Таблиця 2

j	1	2	3	4	5	6	7	8	9	10
p_j	12/25	14/25	4/25	13/25	23/25	7/25	2/25	11/25	12/25	9/25
q_j	13/25	11/25	21/25	12/25	2/25	18/25	23/25	14/25	13/25	16/25
σ_j^2	0,250	0,246	0,134	0,250	0,074	0,202	0,074	0,246	0,250	0,230

2. *Кореляційні зв'язки*. З таблиці 1, наприклад, бачимо, що учасники тестування у завданнях 3 та 7 показали майже однакові результати. Виникає думка про «однаковість» цих двох завдань у певному розумінні. Тобто, можна говорити про певний зв'язок між ними. На перший погляд, існує велика схожість результатів тестування за завданнями 1 і 2. З іншого боку, в поле зору потрапляє і те, що учасники тестування показали майже протилежні результати у завданнях 3 і 5.

Оцінка зв'язків між результатами різних тестувань встановлюється за допомогою *коефіцієнта кореляції*.

Результати i-го та j-го завдань тесту розглядаємо як значення i-ї та j-ї випадкових величин.

Коефіцієнтом кореляції випадкових величин X та Y називається число

$$\rho(X,Y) = \frac{M\left(X \cdot Y\right) - M\left(X\right) \cdot M\left(Y\right)}{\sqrt{D\left(X\right) \cdot D\left(Y\right)}},$$

де M – математичне сподівання, D – дисперсія.

Якщо результати тестування подаються у дихотомічній шкалі, то коефіцієнт кореляції між i-м та j-м завданнями тесту називають φ *- коефіцієнтом кореляції* і обчислюють за формулою

$$\varphi_{ij} = \frac{p_{ij} - p_i \cdot p_j}{\sqrt{p_i q_i \cdot p_j q_j}},$$

де p_{ij} – частка учасників, які успішно впоралися з i-м та j-м завданнями тесту; p_i – частка учасників, які впоралися з i-м завданням ($q_i = 1 - p_i$); p_j – частка учасників, які впоралися з j-м завданням ($q_j = 1 - p_j$).

Знайдемо, наприклад, φ-коефіцієнт кореляції між 3-м та 7-м завданнями тестування, результати якого подано таблицею 1.

$$\varphi_{37} = \frac{p_{37} - p_3 \cdot p_7}{\sqrt{p_3 q_3 \cdot p_7 q_7}} = \frac{(2/25) - (4/25) \cdot (2/25)}{\sqrt{(4/25)(21/25) \cdot (2/25)(23/25)}} = 0,676.$$

Оскільки цей коефіцієнт досить близький до 1, а тестування носить тематичний характер, то можна робити висновок про те, що 3-є та 7-е завдання тесту мають приблизно однакову складність і, отже, в разі потреби їх можна міняти одне на одне.

Аналогічно встановлюємо тісноту зв'язків між усіма завданнями тесту, результати якого подано таблицею 1. Зауважимо лише, що з такої таблиці (дихотомічної матриці) вилучаються не тільки стовпці, що складаються лише з нулів або лише з одиниць, але й такого самого типу рядки. Адже наявність таких рядків чи стовпців не несе нової інформації при проведенні кореляційного аналізу. Таким чином, вилучивши з таблиці 1 сьомого та двадцятого учасників тестування і провівши відповідні обчислення, отримаємо квадратну кореляційну матрицю тестових завдань у вигляді таблиці 3.

Останній рядок таблиці 3 (він є сумою всіх інших рядків) демонструє тісноту кореляційного зв'язку окремого завдання тесту з усім тестом. Звідси бачимо, зокрема, що найменше корелює з усім тестом завдання 5. Це є приводом для подальшого експертного аналізу.

Таблиця 3

$j \setminus j$	1	2	3	4	5	6	7	8	9	10
1	1	0,313	0,405	0,394	0,204	0,422	0,223	0,214	0,128	0,397
2	0,313	1	0,340	0,214	-0,187	0,521	0,187	0,062	0,313	0,272
3	0,405	0,340	1	0,112	0,083	0,652	0,550	0,181	0,146	0,530
4	0,394	0,214	0,112	1	-0,204	0,172	0,204	0,137	0,220	-0,032
5	0,204	-0,187	0,083	-0,204	1	-0,359	0,045	-0,243	-0,223	0,156
6	0,422	0,521	0,652	0,172	-0,359	1	0,443	0,278	0,224	0,190
7	0,223	0,187	0,550	0,204	0,045	0,443	1	-0,187	0,223	0,292
8	0,214	0,062	0,181	0,137	-0,243	0,278	-0,187	1	0,592	0,096
9	0,128	0,313	0,146	0,220	-0,223	0,224	0,223	0,592	1	0,215
10	0,397	0,272	0,530	-0,032	0,156	0,190	0,292	0,096	0,215	1
$\sum \varphi_{ij}$	**3,700**	**3,035**	**3,999**	**2,217**	**0,272**	**3,543**	**2,980**	**2,130**	**2,838**	**3,116**

3. *Валідність тесту*. Для розробника тесту надважливим завданням є створення такого тесту, який повністю відповідатиме поставленій меті тестування. Іншими словами, тест повинен бути *валідним.* Зокрема, якщо метою тестування є розподіл студентів на слабких та сильних, то зрозуміло, що кожне завдання тесту повинно нести свою функцію. Цієї мети не досягти, якщо, наприклад, усі завдання тесту будуть складними або усі завдання тесту будуть легкими. Отже, кожне завдання тесту має бути у певних, наперед закладених зв'язках з іншими завданнями, для забезпечення цієї мети.

Поняття валідності пройшло певний шлях розвитку і є доволі складним. Розглядаються різні види валідності.

На думку американського психолога Анни Анастазі [5], валідність тесту – це «поняття, яке визначає, що вимірює тест і наскільки якісно це здійснюється».

Валідність – це відповідність між тим, що ми робимо і тим, що нам треба зробити.

Тест є невалідним, якщо результати, отримані в ньому, не вимірюють того, що закладалося у меті тестування. Невалідність тесту може мати цілий комплекс причин: це і неякісні тестові завдання, і сама процедура тестування, і методики оцінювання тощо.

У теорії освітніх вимірювань є різні підходи до аналізу якості тестів. Наприклад, дослідники І. Є. Булах та М. Р. Мруга пропонують підхід до аналізу тестів та тестових завдань, який базується на визначенні їхніх головних характеристик та параметрів, виходячи з того, що загальне поняття валідності поділяється за функціональною ознакою на *валідність методу* та *валідність тесту*.

Валідність методу – це валідність змісту, відповідності, прогнозу.

Валідність тесту – це валідність тестових завдань, процедури тестування, процедури оцінювання.

Валідність методу – це відповідність того, що вимірюється цим методом, тому, що він повинен вимірювати. *Валідність відповідності* – це відповідність результатів вимірювання та оцінювання, отриманих різними методами. *Змістова валідність* – це характеристика тесту, що відображає міру впевненості у тому, що завдання тесту досить повно відображають зміст певної сфери знань, точно відображають істотні навички екзаменованого, зміст програми, підручників та мету навчання. Валідність змісту та валідність відповідності вимірюються кількісно через коефіцієнт кореляції, який визначається як коефіцієнт кореляції між результатами тестування та результатами інших вимірювань, виконаних на тій самій групі з того самого предмета іншим методом. Якщо цей коефіцієнт кореляції $\geq 0{,}6$, то результат тестування вважається валідним.

Прогностична валідність – це така характеристика тесту, яка відображає міру впевненості, що отримані за тест оцінки добре прогнозують майбутні досягнення екзаменованої особи [6]. До таких тестів відносяться *тести здібностей* та *тести відбору*. У цьому випадку коефіцієнт валідності обчислюється як

коефіцієнт кореляції між результатами тестування та критерієм, який виражається в оцінці реальної діяльності екзаменованих осіб.

Основними методами обчислення різних показників валідності є методи кореляційного аналізу. Наприклад, для оцінювання валідності різних завдань тесту використовується *коефіцієнт бісеріальної кореляції*. Його використовують тоді, коли один набір значень задано у дихотомічній шкалі, а інший – в інтервальній. Формула для обчислення цього коефіцієнта має вигляд $(r_{bis})_j = \frac{(\bar{X}_1)_j - (\bar{X}_0)_j}{s_X} \cdot \frac{(N_1)_j - (N_0)_j}{uN\sqrt{N^2 - N}}$, де $(\bar{X}_1)_j$ – середнє значення індивідуальних балів тих екзаменованих осіб, які правильно відповіли на j-е завдання тесту, $(\bar{X}_0)_j$ – середнє значення індивідуальних балів тих екзаменованих осіб, які неправильно відповіли на j-е завдання тесту; s_X – стандартне відхилення на множині значень індивідуальних балів; $(N_1)_j$ – число осіб, які правильно відповіли на j-е завдання тесту, $(N_0)_j$ – число осіб, які неправильно відповіли на j-е завдання тесту; $N = (N_1)_j + (N_0)_j$ – число всіх екзаменованих осіб; u – квантиль стандартного нормального розподілу порядку N_1 / N. Часто використовується *точково-бісеріальний коефіцієнт кореляції*:

$$(r_{pbis})_j = \frac{(\bar{X}_1)_j - (\bar{X}_0)_j}{s_X} \cdot \sqrt{\frac{(N_1)_j \cdot (N_0)_j}{N\sqrt{N-1}}}.$$

4. *Надійність тесту***.** Надійність тесту є однією з його фундаментальних характеристик. Якщо порівнювати поняття надійності з поняттям валідності, то поняття надійності є більш технічною характеристикою, пов'язаною з проблемами точності вимірювань в освіті. Якщо не вдаватися до точного означення, то можна сказати, що тест вважається надійним, коли при багаторазовому його використанні у схожих умовах отримуємо однакові результати.

З точки зору порівняння валідності та надійності, треба зазначити, що не надійний тест не може бути валідним, проте не валідний тест може бути надійним. Адже, якщо, наприклад учням другого класу роздати контрольні картки для перевірки їхнього логічного мислення і виділити на цей контроль надто малий проміжок часу, то тест не виконає покладеної на нього функції. Він лише перевірить швидкість читання, залишаючись при цьому надійним за своїм призначенням.

У зв'язку з цим у класичній теорії тестування розроблено класичну модель тестової оцінки, в основі якої лежить твердження про те, що дослідна (тестова) оцінка екзаменованої особи складається з істинної оцінки та похибки її вимірювання. Зрозуміло, що похибки вимірювань пов'язані з поняттям дисперсії. Тому *коефіцієнт надійності*, як числова характеристика тесту, визначається у вигляді відношення дисперсії істинної оцінки до дисперсії дослідної оцінки. Мовою кореляційного аналізу, це є коефіцієнт кореляції між оцінками за цілком паралельні форми того самого тесту.

Оскільки істинна оцінка екзаменованої особи нам не відома, то для знаходження її точкової характеристики або довірчого інтервалу існують емпірично підібрані формули (формули Галліксена, Спірмена-Брауна, Рюлона-Гуттмана, альфа-коефіцієнт Кронбаха та інші) [3,4]. При цьому для оцінки стандартної похибки ще до проведення тестових вимірювань, треба: 1) корінь квадратний із кількості завдань тесту помножити на 0,45, у випадку тесту середньої складності (коли $p = 0,5$); 2) корінь квадратний із кількості завдань тесту помножити на 0,3, у випадку легкого тесту складності $p = 0,9$.

Отже, надійність тестування, це стійкість його результатів при повторних тестуваннях.

Зрозуміло, що надійність тесту залежить від надійності всіх його завдань. *Показником надійності* i-го дихотомічного завдання називається число, що дорівнює $\sqrt{p_i q_i}\rho_{iT}$, де p_i – складність завдання, $q_i = 1 - p_i$, ρ_{iT} – коефіцієнт точково-бісеріальної кореляції між оцінкою за i-е завдання та оцінкою за весь тест.

Для розробника тесту важливо також контролювати стандартну похибку S_ρ для коефіцієнта кореляції. Це можна робити по різному в залежності від того, як оцінювалася дискримінативність тестових завдань. Якщо вона оцінювалася через φ-коефіцієнт кореляції, то використовують просту наближену формулу:

$s_\rho = 1/\sqrt{N-1}$, N – обсяг вибірки, коли $N \geq 50$. Вважається, що критичне значення дорівнює $2s_\rho > 0$. Таким чином у тесті треба залишити ті завдання, для яких коефіцієнт кореляції не менший за $2s_\rho$.

5. *Оцінка якості дистракторів*. У контексті ефективності закритих математичних тестів важливим моментом є якісний добір дистракторів. Числова оцінка правдоподібності дистракторів спирається на знаходження частки тих учасників тестування, які вибрали неправильні відповіді при тестуванні.

Припустимо, що 200 осіб отримали тестове завдання з чотирма варіантами відповіді, серед яких один – правильний, а решта три – дистрактори. Якщо це завдання правильно виконали 70% усіх учасників тестування, то його складність дорівнює $p = 0,7$. Решта 60 учасників вибрали за правильну відповідь один із дистракторів. Якщо дистрактори однаково привабливі, то кожен із них вибрали приблизно по 20 учасників. Іншими словами, вибір кожного із дистракторів підпорядкований майже рівномірному закону їхнього розподілу.

На практиці прийнято вважати, що дистрактор, який вибрали менше ніж 5% учасників від числа тих, які відповіли неправильно, треба вилучити. Це свідчить про дуже нерівномірний розподіл дистракторів за якістю, і отже його залучення до тесту є невдалим.

З іншого боку, рівномірність розподілу дистракторів лише допомагає контролювати їхню правдоподібність і не означає нічого більшого за це.

У наших міркуваннях ми не закладаємо ефекту вгадування. Адже, якщо група осіб, у якій проводиться тестування, схильна до вгадування, то на практиці це спонукає до переоцінки складності тестових завдань. У разі, коли істинна складність тестового завдання дорівнює 0,5, а 50% тих учнів, які не знають правильної відповіді, вгадують її з імовірністю 0,5, то дослідна складність завдання дорівнює 0,75. Крім того, окремі вгадування не є «сліпими», бо правильну відповідь іноді можна вгадувати методом раціонального відбору за певними ознаками. Проте інколи це можна вважати проявом деяких знань. У цьому контексті уже виникає інше цікаве запитання: як за отриманою дослідною складністю тестового завдання знайти істинну складність завдання?

Ці міркування приводять до необхідності побудови якісних дистракторів. Неякісні дистрактори є однією з причин неправильної оцінки істинної складності тестового завдання.

Одним із методів аналізу якості дистракторів у тестових завданнях з однією правильною відповіддю є метод порогових груп [4]. Найважливіша тенденція цього аналізу проявляється у тому, що найсильніші групи екзаменованих осіб найчастіше вибирають правильну відповідь, а всі дистрактори, якщо вони якісні, мають приблизно однакову кількість «голосів».

Цілком валідні завдання тесту повинні мати хорошу роздільну здатність (дискримінативність). Для вимірювання роздільної здатності тестових завдань використовуються різні показники. Найуживаніші з них базуються на понятті кореляції або на методі порогових груп.

У випадку використання дихотомічної шкали найпростішим показником роздільної здатності є *індекс дискримінативності*, який визначається таким чином. За результатами загального тестування всіх його учасників поділяють на 3−4 рівні за їх чисельністю групи. Тоді індексом дискримінативності певного тестового завдання називається число $d = p_{слн} - p_{слб}$, де $p_{слн}$ – частка учасників тестування найсильнішої групи, які правильно відповіли на дане завдання тесту; де $p_{слб}$ – частка учасників тестування найслабкішої групи, які правильно відповіли на дане завдання тесту. Оскільки найбільше значення кожної з цих часток може дорівнювати 1 (коли всі представники групи відповіли правильно), а найменше – може дорівнювати 0 (коли всі представники групи відповіли неправильно), то величина $d = p_{слн} - p_{слб}$ може набувати значень від −1 до 1.

Зважаючи на простоту цього показника, він є дуже прийнятним на рівні студентської групи чи лекційного потоку кількох груп. Якщо $d \geq 0,4$, то це вважається практично задовільним.

Визначення роздільної здатності завдань тесту з використанням поняття коефіцієнта кореляції може здійснюватися через визначення:

- φ-коефіцієнта кореляції;
- точково-бісеріального коефіцієнта кореляції;
- бісеріального коефіцієнта кореляції;
- тетрахоричного коефіцієнта кореляції.

φ-коефіцієнт кореляції є окремою формою коефіцієнта кореляції Пірсона для випадку дихотомічних завдань тесту. Він відображає зв'язок складностей завдань тесту.

Точково-бісеріальний коефіцієнт кореляції також використовують у випадку дихотомічної шкали.

В основу використання бісеріального коефіцієнта кореляції взято припущення про те, що якість завдань розподілена за нормальним законом.

Якщо потрібно вдатися до дихотомізації нормального закону розподілу випадкової величини, то використовується так званий тетрахоричний коефіцієнт кореляції.

Ми не обговорюємо питання про те, в якій ситуації найдоцільнішим є використання того чи іншого показника роздільної здатності. Очевидно, що найпростішим є використання індексу дискримінативності d.

Висновки. У практичній тестології при апробації надійності тесту постає питання про визначення обсягу вибірки екзаменованих осіб. Правила, для визначення мінімального розміру такої вибірки, не існує. Зрозуміло, що чим більший обсяг вибірки, тим надійнішими будуть отримані на основі тестування числові характеристики параметрів тестових завдань. Наприклад, для проведення апробації тестів регіонального рівня бажано, щоб обсяг вибірки містив у собі не менше як кілька сотень осіб. Проте існує емпіричне правило: обсяг вибірки повинен перевищувати кількість завдань у тесті не менш як у 5 разів.

Якщо перед конструюванням тесту у його розробника в наявності занадто багато завдань, то перед ним постає проблема раціонального їх відбору. У такому разі треба до тесту залучати одне за одним ті завдання, які є найвагомішими для забезпечення потрібної надійності та валідності. До того ж, під час розробки тесту заданої якості кожен розробник намагатиметься залучити до нього оптимальну кількість завдань, з метою економії часу та інших ресурсів.

Якщо сукупність завдань не дозволяє їх великого перебирання і внесок кожного з них у тест є позитивним, то до тесту насамперед треба внести ті завдання, які добре корелюють з усім тестом (критерієм).

У випадку, коли для роздільної здатності завдань використовувався бісеріальний коефіцієнт кореляції, стандартну похибку оцінюють за формулою (див. [4]) $s_{bis} = \frac{\sqrt{pq/N-1}}{Y}$. Тут також використовується стандартний нормальний розподіл випадкової величини. Стандартна похибка бісеріальної кореляції є незначною для тестових завдань середньої складності, але вона збільшується для завдань, складність яких наближається до мінімального чи максимального рівнів.

У контексті питань, які обговорюються нами, важливим є питання про вплив складності завдань на їх відбирання до тесту. Для тесту, головною метою якого є максимальна диференціація екзаменованих осіб, більшість завдань повинна бути середньої складності. І лише незначна їх кількість повинна охоплювати легкі та складні завдання. Якщо в результаті тестування серед усіх інших треба оцінити осіб з екстремальними здібностями, то такий тест повинен містити також завдання екстремальної складності. У такому разі перед розробником постає завдання валідизації тесту.

Список використаних джерел:

1. Найко Д. А. Особливості побудови тестових завдань з математичних дисциплін / Д. А. Найко // Сучасні інформаційні технології та інноваційні методи навчання у підготовці фахівців: методологія, теорія, досвід, проблеми. //Зб. наук. пр.– Випуск 48 / редкол. – Київ-Вінниця: ФОП Тарнашинський О. В., 2017. – С. 154 – 160.

2. Краєвська О. Д. Визначення однорідності груп при дослідженні процесу формування комунікативної компетентності майбутніх менеджерів-аграріїв. / О. Д. Краєвська, Д. А. Найко // Економіка. Фінанси. Менеджмент: актуальні питання науки і практики. / Зб. наук. праць Вінницького національного аграрного університету. – 2015. – №1. – С. 70 − 81.

3. Вимірювання в освіті: Підручник / За редакцією О. В. Авраменко. − Кіровоград: «КОД», 2011. – 360 с.

4. Ковальчук Ю. О. Теорія освітніх вимірювань. – Ніжин: Видавець ПП Лисенко М. М., 2012. – 200 с.

5. Конструювання тестів. Курс лекцій: навч. посіб. /Л. О. Кухар, В. П. Сергієнко. – Луцьк, 2010. – 182 с.

6. Шевчук О.Ф. Наскрізний електронний посібник фахового спрямування: переваги та особливості **/** О.Ф. Шевчук **//** Сучасні інформаційні технології та інноваційні методики навчання у підготовці фахівців: методологія, теорія, досвід, проблеми // Зб. наук. пр. – Випуск 48 / редкол. – Київ-Вінниця: ФОП Тарнашинський О. В., 2017. – С. 192 − 194.