УДК 378.4

О.М. КУТИК

ОЦІНЮВАННЯ ЯКОСТІ ПЕДАГОГІЧНИХ ТЕСТІВ ЯК НАУКОВО-МЕТОДИЧНА ПРОБЛЕМА

Статтю присвячено питанню оцінювання якості педагогічного тесту. Дано загальну характеристику основних етапів створення педагогічного тесту.

Ключові слова: оцінювання якості педагогічного тесту, індекс дискримінації, індекс складності, асиметрія, ексцес, валідність тесту.

Прагнення української середньої освіти до підвищення її якості та запровадження ЗНО зробили досить актуальною проблему оцінювання знань учнів. Серед методів оцінювання знань найбільшої популярності останнього часу набуло тестування. Незважаючи на значну кількість публікацій у науково-педагогічних періодичних виданнях із питань застосування тестів, на практиці досі залишається не вирішеними питання якості тестових матеріалів, які використовуються в навчальному процесі. Однією з головних причин цього є недостатній рівень теоретичних знань педагогічних працівників загальноосвітніх навчальних закладів з основ тестології, що не дає їм можливості оцінити їх якість.

На сьогодні у педагогічній теорії достатньо широко представлені дослідження з різних аспектів тестолоії. Найбільш відомими є праці В. Аванесова [1], А. Анастазі [2], І. Булах [3], В. Кіма [4], О. Майорова [6], Л. Паращенко [8], М. Челишкової [9] та ін. Водночас, значний теоретичний матеріал, який викладено у працях наведених авторів, не дає змоги багатьом педагогічним працівникам скласти алгоритм дій з оцінювання якості тестів, що значно ускладнює розуміння ними зазначеної проблеми.

Мета статті полягає в загальній характеристиці та визначенні алгоритму процесу створення та оцінювання якості педагогічного тесту.

Перш за все, необхідно визначити термінологічне поле теми. Досить часто на практиці відбувається підміна таких понять, як: контроль, педагогічне вимірювання, оцінювання.

Педагогічним контролем називають єдину дидактичну та методичну систему перевірної діяльності, спрямованої на виявлення результатів навчального процесу й підвищення його ефективності [8, с. 7]. Як відомо, контроль розглядають як одну із функцій управління, спрямовану на вирішення трьох завдань: виявлення відхилень фактичних результатів управління від передбачених, з'ясування причин розбіжності мети та результатів управління, визначення змісту регулювальної діяльності з метою зведення до мінімуму наявних відхилень.

Педагогічне вимірювання, як зазначає В. Кім, є процесом встановлення відповідності між оцінюваними характеристиками учнів і точками емпіричної шкали, в яких відносини між різними оцінками характеристик виражені властивостями числового ряду [4, с. 17]. Тоді як *оцінювання (evaluation)* розглядається як процес формулювання висновків на основі порівняння кількісних показників, отриманих із різних джерел, зі стандартами та здійснюється, як правило, для вдосконалення програм, зокрема

програм навчання. Оцінювання поділяється на формувальне та підсумкове; оцінювання результатів і процесів [7, с. 11]. Таким чином, контроль, вимірювання та оцінювання не є тотожними поняттями. Вони відрізняються як за своєю суттю, так і функціями, які виконують.

Розробка проблеми вимірювань вимагає вирішення трьох взаємопов'язаних питань: навіщо, що та чим вимірювати? Відповідь на перше запитання є доволі простою. Вона безпосередньо пов'язана з постановкою цілей контролю. Якщо мета – оцінювання досягнень учнів, то головна увага приділяється перевірці та виявленню обсягу засвоєних знань або вмінь. При оцінюванні досягнення учнів як предмета вимірювання (другого питання – що?) зазвичай виділяють рівень та якість підготовки. Для контролю знань учнів досить широко використовується такий інструмент, як тести.

Тести – це *система* тестових завдань різної складності, які дають змогу якісно та ефективно виміряти рівень і структуру підготовки учнів [9, с. 19]. Необхідно звернути увагу на словосполучення "система завдань". Вилучення одного або декількох тестових завдань із тесту призведе до необ'єктивності вимірювання навчальних досягнень учнів. Тест має відповідати певним характеристикам: бути валідним (це відповідність того, що вимірюють цим методом, тому, що він має вимірювати), точним (мінімальна або систематична похибка, з якою можна провести вимірювання цим методом), надійним (точність діагностичних вимірювань, а також стабільність і стійкість їх результатів щодо впливу різних зовнішніх факторів), а також об'єктивним (характеристика якості методу, яка виявляється у тому, наскільки отримані результати відповідають дійсності, є достовірними).

У системі тестування існує два види оцінювання (відповідно, і тестів): нормативно орієнтоване та критеріально орієнтоване. Нормативно орієнтоване оцінювання дає змогу ранжувати тих, кого випробують, за рівнем знань. Таким чином, результати тестування за нормативно орієнтованим тестом дають можливість порівняти навчальні досягнення екзаменованих один з одним. Якість такого тесту (як інструменту вимірювання) визначається за допомогою статистичних методів аналізу результатів його апробації.

Критеріально орієнтований тест являє собою систему завдань, які дають змогу виміряти рівень навчальних досягнень щодо повного обсягу знань, умінь і навичок, які повинні бути засвоєні учнями. Цей тест дає оцінки лише за дихотомічною шкалою: виконав – не виконав, пройшов – не пройшов, залік – не залік. Визначити якість такого тесту досить важко [6, с. 38].

Розробка тесту проходить довгим та складним шляхом. О. Майоров наводить схему етапів складання тестів навчальних досягнень для різного рівня використання (табл. 1) [6, с. 50].

Етап постановки мети використання тесту є визначальним. Від результатів його виконання залежить якість тесту та відповідність його змісту тому, що передбачено виміряти.

У процесі формулювання мети розробнику, насамперед, необхідно вирішити питання про те, які результати навчальної діяльності учнів він хоче оцінити за допомогою тесту.

Таблиця 1

Схема етапів складання тестів навчальних досягнень для різного рівня використання

№ з/п	Етапи складання тесту	Рівень використання		
		Використовується педагогом для професійних потреб	Використовується для внутрішніх потреб школи	Використовується для підсумкової атестації учнів, найбільше адміністративне використання
1	Визначення цілей тестування	Так	Так	Так
2	Відбір змісту навчального матеріалу	Так	Так	Так
3	Конструювання технологічної матриці	Бажано	Так	Так
4	Складання тестових завдань	Так	Так	Так
5	Побудова вибірки та апробація завдань та тесту		Так	Так
6	Компоновка завдань для апробації	Так	Так	Так
7	Апробація тестових завдань	Так	Так	Так
8	Визначення та розрахунок показників якості тестових завдань	Бажано	Бажано	Так
9	Відбраківка завдань та складання тесту	Так	Так	Так
10	Апробація тесту	Бажано	Бажано	Так
11	Визначення та розрахунок показників якості тесту	Бажано	Бажано	Так
12	Складання остаточного варіанту тесту	Так	Так	Так

У тестології прийнято застосовувати класифікацію рівнів засвоєння знань (таксономія когнітивної сфери), розроблену Б. Блумом: I рівень – знання (здатність запам’ятовувати та відтворювати вивчений матеріал); II рівень – розуміння (здатність інтерпретувати навчальний матеріал, перетворювати його з однієї форми подання на іншу); III рівень – застосування (можливість використовувати засвоєний матеріал у нових ситуаціях та умовах); IV рівень – аналіз (здатність поділяти та структурувати навчальний матеріал таким чином, що стає зрозумілою загальна організаційна структура); V рівень – синтез (здатність комбінувати елементи для отримання нового цілого); VI рівень – оцінювання (здатність оцінювати значення того чи іншого матеріалу) [5, с. 39].

Під час планування тесту розробнику доводиться думати про те, що далеко не весь набір цілей, а відповідно, і навчальних елементів можна відобразити в змісті тесту. У зв'язку із цим, їх набір необхідно структурувати так, щоб у тест потрапили найважливіші.

Із цією метою розробляють специфікацію (матрицю) тесту, яка має запобігти незбалансованості та диспропорції охоплених тестом тем курсу або дидактико-психологічних цілей. При проведенні специфікації максимально чітко визначають навчальні цілі відповідного навчального предмета та теми, за якими складається тест, визначають їх основний зміст, будують таблицю, у якій відбивається ця інформація та зазначається кількість тестових завдань по кожному елементу.

Існують і правила оформлення тесту, який складено відповідно до його специфікації. Так, складність завдань у тесті має збільшуватися зі зростанням порядкового номера. Завдання окремих форматів розташовуються групами. До кожної групи завдань має бути наведена інструкція щодо їх виконання. Кількість форматів завдань, використаних в одному тесті, не повинна перевищувати трьох.

Чи відповідає розроблений тест зазначеним вище характеристикам, визначається шляхом його апробації та статистичного аналізу отриманих результатів. З якою метою здійснюють аналіз і чому це надзвичайно важливий етап технологічного циклу стандартизованого тестування, без якого не можна створити тест вирішальної значущості для визначальних адміністративних висновків? Це пояснюється двома причинами.

По-перше, доки ми не впевнимося, що інструмент вимірювання є якісним, ми не можемо робити висновки щодо отриманих результатів.

По-друге, ми повинні проаналізувати його елементи – тестові завдання – та вилучити неякісні завдання, якщо такі є.

Під час аналізу якості інструменту вимірювання використовують аналітичні (експертні) та емпіричні (статистичні, математичні) методи.

Сукупність емпіричних методів визначення якості інструменту вимірювання, що базуються на даних самого вимірювання, називають психометричним аналізом, до якого відносять: середнє (M), мінімальне та максимальне значення; середнє квадратичне відхилення (σ); коефіцієнт надійності R (α-Кронбаха або Кьюдер-Ричардсона); помилка вимірювання – E; коефіцієнт валідності (якщо наявний); коефіцієнт асиметрії – *As* (асиметричність кривої); коефіцієнт ексцесу – *Ex* (зміна гостроти кривої у вершині).

Алгоритм аналізу тестових завдань і тесту в цілому може складатися з таких елементів:

1. Побудова й аналіз матриці результатів.
2. Визначення міри складності та дискримінативності тестових завдань.
3. Визначення варіації балів (різниця максимально набраного та мінімально набраного балів).
4. Визначення мір центральної тенденції.
5. Побудова гістограми та полігонів частот розподілів балів.
6. Визначення надійності тесту.

При цьому послідовність проведення аналізу може бути різною [1–3; 6; 9]. Класичними психометричними характеристиками тестового завдання є складність тестового завдання (індекс складності P) та розподільча здатність тестового завдання (індекс дискримінативності ID, point biserial).

Що таке складність тестового завдання? Те, що є складним для учня першого класу, не є складним для одинадцятикласника. Тобто тестове завдання є складним лише щодо конкретної групи учнів, конкретного рівня, класу, школи тощо. Таким чином, індекс складності встановлює, наскільки конкретне тестове завдання є складним для осіб, яких тестували. Його визначають за відсотком учнів, що правильно відповіли на це тестове завдання, і він може змінюватися від 0 до +1,0. Складність (P-value) визначається за формулою:

$$P_i = \frac{R_i}{N}, \tag{1}$$

де R_i – кількість учнів, які виконали завдання правильно;

N – загальна кількість учнів.

Тобто, чим вищий є показник, тим легше тестове завдання: якщо 100% учнів відповіли правильно, то отримуємо показник 1. Залежно від значення складності тестове завдання класифікується як складне або легке таким чином (міжнародна практика): 0,20 < P < 0,36 – надто складне; 0,36 < P < 0,84 – середньої складності; 0,84 < P – надто легке [3, с. 24].

Український центр оцінки якості освіти використовує такі діапазони коефіцієнтів складності тестового завдання (табл. 2).

Таблиця 2

Діапазони рівнів складності тестових завдань

Інтервал індексу складності	Характеристика завдань
Понад 0,8	Дуже легкі
0,60–0,79	Легкі
0,40–0,59	Оптимальні
0,20–0,39	Складні
0,19 і нижче	Дуже складні

Оптимальним вважається тест, який містить 20% легких завдань, 60% завдань середньої складності та 20% складних і дуже складних завдань. Якщо в тесті буде збільшено кількість складних завдань за рахунок легких і середнього рівня складності, результати вимірювання будуть заниженими, і навпаки.

Розподільча здатність тестового завдання – індекс дискримінативності – вказує, наскільки добре це тестове завдання розрізняє екзаменованих із високим балом і екзаменованих із низьким балом. Індекс може коливатися від –1 до +1. У випадку, коли індекс дорівнює “0”, всі екзаменовані відповіли однаково (добре або погано). Розподільчу здатність можна розрахувати як коефіцієнт кореляції між балом за весь тест і балом за конкретне тестове завдання.

Найбільш простий метод розрахунку (“ручний”) полягає у встановленні різниці між складністю тестового завдання для групи сильних (Hi) і групи слабких (Lo) учнів: ID = P diff (Hi) – P diff (Lo). У разі ID < 0,0 – сильні учні відповідають гірше за слабких, у разі ID < 0,2 – тестове завдання недостатньо (або погано) розподіляє учнів. Тестове завдання має достатню розподільчу здатність у разі ID > 0,2 [3, с. 25].

Наріжним каменем тестології є припущення про нормальний розподіл показників, за якими оцінюють здібності людей до навчання. Чим це зумовлюється? Ураховуючи, що розподіл ознак у живій природі підпорядковується закону нормального розподілу (закону Гауса), доцільно припустити, що й знання, уміння та навички у вибраній групі з достатньою кількістю учнів також розподіляються відповідно до цього самого закону. Функція (графік) нормального розподілу має вигляд дзвону, тобто 70% значень знаходяться в центрі, а інші сходять нанівець до країв розподілу.

Якщо дані апробації тесту відповідають закону нормального розподілу, то це свідчить про якість тесту. Постає запитання: за яких умов ми можемо вважати, що розподіл даних відповідає закону Гауса? Перший крок – необхідно визначити міри центральної тенденції (середнє, мода, медіа-на).

Мода – це таке значення, яке найчастіше зустрічається серед результатів виконання тесту. Середнє арифметичне визначають підсумовуванням усіх значень сукупності та подальшим діленням на їхню кількість. Медіана – це величина ознаки, яка лежить всередині упорядкованого за зростанням або спаданням ряду розподілу та ділить його на дві рівні частини.

При нормальному розподілі результатів показники моди, медіа-ни та середнього арифметичного однакові або дуже близькі одне до одного.

За умови, що дані результатів розподілені відповідно до цього закону, ми можемо вважати їх *достовірними,* коли 70% учнів правильно впоралися із кількістю завдань, які знаходяться в діапазоні від 30–70%. Для цього необхідно визначити міру розпорошення – дисперсію та стандартне відхилення.

Дисперсія – середня арифметична із квадратів відхилень, варіант від їх середньої арифметичної. Дисперсія змінюється від нуля до безкінечності. Крайнє значення “0” означає відсутність мінливості, коли значення змінної постійне. Стандартне відхилення (середнє квадратичне відхилення) обчислюється як корінь квадратний із дисперсії. Воно показує, що 70% даних знаходиться в діапазоні між: середнє значення мінус стандартне відхилення та середнє значення плюс стандартне відхилення [9, с. 232].

Інший спосіб встановлення якості інструментарію пропонує М. Челишкова. Вона вважає, якщо середнє арифметичне приблизно дорівнює потрійному значенню стандартного відхилення, то є підстави вважати дисперсію оптимальною, а розподіл тестових балів близьким до нормального. При цьому тест буде мати достатню розподільчу здатність [9, с. 238]. Відзначимо, що це твердження справедливе не для всіх випадків. Можливі ситуації, коли середнє арифметичне значно більше за потрійне стандартне відхилення, але розподіл тестових балів, проте достатньо близьке до нормального.

Наступний крок – необхідно порівняти показники асиметрії й ексцесу.

Асиметрія – це властивість розподілу вибірки, яка характеризує несиметричність розподілу відносно середнього балу. Ексцес – це міра крутизни кривої розподілу. Асиметрія й ексцес бувають позитивними і негативними. Ці моменти складають набір особливостей розподілу при аналізі даних.

Для нормального розподілу As=0, Ex=0. Як свідчить досвід, коефіцієнти асиметрії й ексцесу ніколи не дорівнюють 0. Тому ці коефіцієнти повинні бути близькі до 0. Загальну характеристику тесту надає порівняння показників асиметрії й ексцесу (табл. 3) [8, с. 40].

Таблиця 3

Аналіз тестів на валідність

As	Ex	Характеристика
= 0	= 0	Тест валідний
< 0	> 0	Тест легкий з низькою розподільчою здатністю
< 0	< 0	Тест легкий з високою розподільчою здатністю
< 0	= 0	Тест легкий
> 0	< 0	Тест складний з високою розподільчою здатністю
> 0	> 0	Тест складний з низькою розподільчою здатністю
> 0	= 0	Тест складний
= 0	> 0	Неоднорідний тест (наявність важких і легких завдань), з низькою розподільчою здатністю
= 0	< 0	Неоднорідний тест (наявність важких і легких завдань), з високою розподільчою здатністю

У дослідженні І. Булах запропоновано схему аналізу валідності тестових завдань, що ґрунтується на результатах статистичної обробки даних тестування з побудовою кривої розподілу кількостей правильних відповідей на конкретне тестове завдання. Невалідними вважаються завдання, на які під час тестування отримано: правильні відповіді більше ніж у 84% осіб, що тестувалися; правильні відповіді менше ніж у 16% осіб, що тестувалися [8, с. 39]. На думку В. Аванесова, невалідними є тестові завдання, які виконали всі учні, або не виконав ніхто [1].

Для проведення розрахунків наведених показників радимо користуватися можливостями комп'ютерної програми Microsoft Excel для Windows. Для цього необхідно, по-перше, створити матрицю результатів; по-друге, застосувати "пакет аналізу" (Сервіс (Дані) / Аналіз даних / Описова статистика).

Оцінювання за результатами тестування вважається адекватним у разі використання надійного та валідного інструментарію.

Валідність тесту показує, наскільки добре тест робить те, для чого він був створений. Визначити коефіцієнт валідності тесту – визначити, як результати виконання тесту співвідносяться з іншими незалежно виставленими оцінками знань екзаменованих. Для визначення валідності потрібно застосовувати незалежний зовнішній критерій, тобто оцінку експерта (вчителя). За коефіцієнт валідності беруть коефіцієнт кореляції результатів тес-

тових вимірювань і критерію. Якщо експертна оцінка знань екзаменованих, отримана незалежно від процедури тестування, представлена числовою послідовністю y_1, , ..., y_n, то коефіцієнт валідності тесту може бути розрахований за формулою [9, с. 344]:

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

де Y_i – послідовність індивідуальних результатів тестування екзаменованих;

y_i – послідовність відповідних індивідуальних оцінок експертів тих самих екзаменованих;

$\overline{Y}$, $\overline{y}$ – середні арифметичні результатів тестування та експертних оцінок;

S_Y, S_y – стандартні відхилення цих оцінок.

Ще одним важливим критерієм адекватності тесту є надійність його результатів, тобто точність вимірювання рівня навчальних досягнень. Надійність тесту визначається перевіркою його внутрішньої узгодженості.

Показником надійності є коефіцієнт α-Кронбаха, що визначається за формулою [6, с. 173]:

$$\alpha = \frac{\hat{e}}{\hat{e}+1}\left(1 - \frac{\sum \delta_i^2}{\delta_y^2}\right), \quad (3)$$

де κ – кількість завдань;

$\sum \delta_i^2$ – сума квадратів стандартних відхилень для завдань;

δ_i^2 – квадрат стандартного відхилення для всього тесту.

Інтервали значень і характеристики тесту за коефіцієнтом надійності показано в табл. 4.

Таблиця 4

Характеристика тесту за коефіцієнтом надійності

Інтервали коефіцієнта надійності	Характеристика тесту
0,90–0,99	Відмінний
0,85–0,89	Дуже хороший
0,80–0,84	Хороший
0,75–0,79	Задовільний
0,70–0,74	Мало задовільний

Висновки. У статті показано, що використання тестування як технології оцінювання знань не є тривіальним завданням і не зводиться лише до створення питань і ймовірних варіантів відповідей. Це тривалий процес, який іноді не приводить до бажаного результату, якщо немає розуміння

всіх тонкощів його проведення, якщо відсутнє володіння основами теорії тестування.

Водночас слід зауважити, що із впровадженням зовнішнього незалежного оцінювання та підготовкою старшокласників до участі у ньому виникає необхідність використання в навчальному процесі середньої школи тестів (а не сукупності окремих завдань у тестовій формі), які б відповідали зазначеним у статті характеристикам. Виходячи із цього вбачаємо за необхідне розробникам тестів публікувати апробований інструментарій з результатами психометричного аналізу. Це дасть можливість учителям бути впевненими в якості тестів, якими вони користуються під час оцінювання навчальних досягнень знань учнів, а відповідно, і в об'єктивності оцінок, які отримують учні.

Список використаної літератури

1. Аванесов В.С. Методологические и теоретические основы тестового педагогического контроля [Электронный ресурс] / Вадим Сергеевич Аванесов // Сайт научно-методической поддержки слушателей курсов автора по теме "Педагогические измерения" – Электронные данные. – Режим доступа: http://testolog.narod.ru.

2. Анастази А. Психологическое тестирование : [пер. с англ.] : в 2 кн. / А. Анастази. – М. : Педагогика, 1982. – Кн. 1. – 316 с.

3. Булах І.Є. Створюємо якісний тест : навч. посіб. / І.Є. Булах, М.Р. Мруга. – К. : Майстер-клас, 2006. – 160 с.

4. Ким В.С. Тестирование учебных достижений : монография / В.С. Ким. – Уссурийск : Изд-во УГПИ, 2007. – 214 с.

5. Лукіна Т.О. Педагогічна діагностика (Модуль 10) : навч.-метод. матеріали / Т.О. Лукіна. – К. : Міністерство освіти і науки України, Проект Світового банку "Рівний доступ до якісної освіти", 2008. – 59 с.

6. Майоров А.Н. Теория и практика создания тестов для системы образования / Алексей Николаевич Майоров. – М. : Интеллект-центр, 2001. – 296 с.

7. Основи педагогічного оцінювання (навчально-методичні та інформаційно-довідникові матеріали для педагогічних працівників). – К. : Майстер-клас, 2005. – Ч. 1. Теорія. – 95 с.

8. Паращенко Л. Тестування учнів у середній школі / Л. Паращенко, В. Леонський, Г. Леонська. – К. : Шкільний світ, 2009. – 128 с.

9. Челышкова М.Б. Теория и практика конструирования педагогических тестов : учеб. пособ. / Марина Борисовна Челышкова. – М. : Логос, 2002. – 432 с.

Кутик А.Н. Оценивание качества педагогических тестов как научно-методическая проблема

Статья посвящена вопросу оценки качества педагогического теста. Дана общая характеристика основных этапов создания педагогического теста.

Ключевые слова: оценка качества педагогического теста, индекс дискриминации, индекс сложности, ассиметрия, ексцес, валидность теста.

Kutik O. The evaluation of quality of pedagogical test as a scientific and methodical problem

The article is devoted to the question of the evaluation of quality of the pedagogical test. The general description of the steps to create of the pedagogical test is shown.

Key words: evaluation of quality of pedagogical test, ID – point beserial, P-value, As – asymmetry, Ex – ekstses, validity of the test.