

STATISTICAL METHODS FOR ASSESSING THE VALIDITY AND RELIABILITY OF MASS INDEPENDENT TESTING ON THE EXAMPLE OF THE ANALYSIS OF THE UEE-2024 RESULTS

*Yanina Honcharenko^{1,2,3}, Nataliia Dyvliash^{2,4}, Olena Koliadenko^{2,5},
Oleksandr Kozlenko^{2,6}*

СТАТИСТИЧНІ МЕТОДИ ОЦІНКИ ВАЛІДНОСТІ ТА НАДІЙНОСТІ МАСОВИХ НЕЗАЛЕЖНИХ ТЕСТУВАНЬ НА ПРИКЛАДІ АНАЛІЗУ РЕЗУЛЬТАТІВ ЄВІ-2024

*Яніна Гончаренко, Наталія Дивляш, Олена Коляденко,
Олександр Козленко*

Abstract. The article substantiates possibility and expediency of applying various statistical methods for analyzing the characteristics of the tests that are used for mass independent evaluation. Different approaches to checking of the hypothesis regarding the normality of the distribution of test results and the interpretation of the obtained results are studied; the recommendations are given for testing the normality of the results of the unified entrance exam (UEE) for the second and third levels of higher education in Ukraine. Some aspects of statistical estimates of the test's reliability indicators are explained, it is also stated with which additional restrictions Cronbach's α is an unbiased estimate indicator of the reliability. Based on the results of one of the UEE sessions, the method of specifying the boundaries of the intervals in which the observed parameters of the empirical assessment of the reliability indicator can change is justified and demonstrated. On the example of the results of the test of general academic competence as a component of UEE, the practical application of the reliability analysis of the test as a whole and its individual blocks and tasks is demonstrated, the obtained results are interpreted. A psychometric and substantive analysis of tasks was conducted, for which the results of the reliability study revealed a potential possibility of a negative influence on the reliability of the test.

Keywords: test reliability, validity of test results, Kolmogorov–Smirnov–Lilliefors test, Shapiro–Wilk test, Cronbach test, psychometric characteristics of the test task, test of general academic competence

¹ Dragomanov Ukrainian State University, Kyiv, Ukraine

² Ukrainian Center for Educational Quality Assessment

³ ya.v.honcharenko@udu.edu.ua, <https://orcid.org/0009-0009-2541-7631>

⁴ dyvljashnatalij@gmail.com, <https://orcid.org/0009-0008-8224-8785>

⁵ lekoliad@gmail.com, <https://orcid.org/0000-0003-3453-2219>

⁶ kozlenkoa@gmail.com, <https://orcid.org/0000-0003-3271-9231>

Анотація. Досліджено властивості й обґрунтовано можливість і доцільність застосування різних статистичних методів для оцінки характеристик тестів, які використовують для масових випробувань. Проаналізовано різні підходи до перевірки ²гіпотези щодо нормальності розподілу результатів тестування та інтерпретації отриманих результатів, сформульовано рекомендації щодо перевірки нормальності результатів єдиного вступного іспиту (ЄВІ) на здобуття другого та третього рівнів вищої освіти в Україні. Обґрунтовано деякі властивості статистичних оцінок показників надійності тесту, встановлено за яких додаткових обмежень показник α Кронбаха є незміщеною оцінкою показника надійності. Обґрунтовано й продемонстровано на результатах однієї із сесій ЄВІ метод уточнення меж проміжків, у яких може змінюватися спостережуване значення емпіричної оцінки показника надійності. на прикладі результатів тесту загальної навчальної компетентності як складової ЄВІ продемонстровано практичне застосування аналізу надійності тесту загалом та окремих його блоків і завдань, інтерпретовано отримані результати. Проведено психометричний і змістовний аналіз завдань, для яких за результатами дослідження надійності було виявлено потенційну можливість негативного впливу на надійність тесту.

Ключові слова: надійність тесту, валідність результатів тестування, тест Колмогорова–Смірнова–Ліллєфорса, тест Шапіро–Вілка, α Кронбаха, психометричні характеристики тестового завдання, тест загальної навчальної компетентності

Вступ

Закон України «Про вищу освіту» [1] визначає перехід до компетентнісноцентричної вищої освіти, на протипагу попередній, знанневоцентричній. Згідно зі ст. 1, п. 1 Закону: «Компетентність — здатність особи успішно соціалізуватися, навчатися, провадити професійну діяльність, яка виникає на основі динамічної комбінації знань, умінь, навичок, способів мислення, поглядів, цінностей, інших особистих якостей».

В освітніх стандартах зафіксовано переліки компетентностей випускників, з яких інтегральна компетентність є інваріантною, а загальні компетентності можуть відрізнятися в різних стандартах. У [2] визначено рекомендований перелік загальних компетентностей, які мають обов'язково формуватись на відповідних рівнях вищої освіти. Ці загальні компетентності відображено у Стандартах вищої освіти всіх спеціальностей. Серед основних загальних компетентностей, якими має володіти випускник першого (бакалаврського) рівня вищої освіти, такі:

- ЗК01 Здатність до абстрактного мислення, аналізу та синтезу,
- ЗК02 Здатність застосовувати знання у практичних ситуаціях,
- ЗК05 Здатність спілкуватись державною мовою як усно, так і письмово.
- ЗК10 Здатність до пошуку, оброблення та аналізу інформації з різних джерел.

Отже, формування обов'язкових загальних компетентностей передбачене стандартами вищої освіти України й не залежить від галузі знань і спеціальності, за якою здобувачі отримують вищу освіту.

Тому при вступі на магістерський рівень вищої освіти з переліків загальних компетентностей виокремлюють ключові, універсальні й найбільш

актуальні для успішності вступника в подальшій навчальній діяльності. Це здатності до аналітичного, логічного й критичного мислення, які є складовими загальної навчальної компетентності (далі — ЗНК). «Критичне мислення» визначає здатність особи осмислювати великі обсяги інформації, поданої в різних формах, структурувати її та робити з неї логічні висновки, порівнювати зміст різних джерел інформації, знаходити в них спільні елементи, аналогії й розбіжності тощо; *аналітичне мислення* визначає здатність особи будувати формальні моделі реальних життєвих ситуацій і вміти досліджувати їх на основі дедуктивних міркувань; *логічне мислення* визначає здатність особи аналізувати судження, міркування, аргументації, знаходити в них вади, визначати, як додаткова інформація може посилювати або послаблювати аргументацію на основі індуктивних і дедуктивних міркувань, тощо» [3].

Тест загальної навчальної компетентності (далі — ТЗНК) спрямований на оцінювання рівня сформованості в здобувачів загальних компетентностей, визначених стандартами вищої освіти. Так оцінювання ЗК01, ЗК02 і ЗК10 відбувається переважно в межах логіко-аналітичного блоку ТЗНК, а ЗК05 і ЗК10 — вербально-комунікативного.

Тест загальної навчальної компетентності є різновидом *ability test*, тестів здатностей (здібностей), які протиставлені «знанцевим» тестам, тестам досягнень, *achievement tests*, або предметним тестам, *subject tests*. Тести здатностей використовують у багатьох країнах для відбору вступників до закладів вищої освіти на бакалаврські (Грузія, Ізраїль, США, Австралія, Швеція тощо) й магістерські програми. Найбільш відомими є тести, розроблені в США: GRE — універсальний вступний тест загальних здібностей для вступу на будь-яку магістерську програму, створений ETS (www.ets.org), і LSAT — спеціалізований вступний тест на магістерські програми правничого спрямування, розроблений LSAC (www.lsac.org) [4].

В Україні впровадження ТЗНК відбувалося в декілька етапів. У 2009–2011 рр. було проведено педагогічний експеримент щодо впровадження ТЗНК для вступників до вищих навчальних закладів, зокрема створення та апробація добірок завдань до вербально-комунікативного й логіко-математичного типів мислення. У 2017 року Міністерство освіти та науки спільно з вищими навчальними закладами правничого спрямування прийняли рішення, що іспит у формі зовнішнього незалежного оцінювання стане обов'язковим для всіх осіб, які вступають на навчання за магістерською програмою за спеціальністю 081 «Право». Єдине фахове вступне випробування (ЄФВВ) для осіб, які виявили бажання взяти участь у конкурсному відборі на навчання для здобуття ступеня магістра за спеціальностями 081 «Право» та 293 «Міжнародне право» складалося з блоків тесту загальної навчальної правничої компетентності й тесту з фахової дисципліни — «Право». У 2022 році для вступу на магістерські програми за спеціальностями галузей знань 05 «Соціальні та поведінкові науки», 06 «Журналістика», 07 «Управління та адміністрування», 28 «Публічне управління та адміністрування», 29 «Міжнародні відносини» було проведено іспити із застосуванням магістерського тесту навчальної компетентності. У 2023 році відповідно до Порядку прийому на навчання для здобуття вищої освіти, затвердженого наказом Міністерства освіти і науки Украї-

ни №276 від 15 березня 2023 року, для відбору на навчання вступників на основі НРК6 або НРК7 для здобуття ступеня магістра, УЦОЯО вперше провів ЄВІ з використанням двох інструментів (ТЗНК й тест з іноземної мови) та ЄФВВ. Того ж року використання ТЗНК для відбору абітурієнтів для вступу на магістерські програми стало постійним, а в 2024 році цей інструмент було також застосовано як допуск до вступу в аспірантуру.

Огляд останніх джерел досліджень і публікацій

Загальні результати проведення вступних випробувань до магістратури й аспірантури у 2024 році представлені в офіційному звіті УЦОЯО [7, 8].

Зміст тесту визначено Програмою тесту загальної навчальної компетентності єдиного фахового вступного випробування для вступу на навчання для здобуття ступеня магістра, затвердженою наказом Міністерства освіти і науки України №158 від 11 лютого 2022 року [5]. Загальна кількість завдань тесту — 27, на виконання завдань тесту відведено орієнтовно 75 хвилин [6].

Тест складається з двох компонентів, вербально-комунікативного (12 завдань) і логіко-аналітичного (15 завдань), і містить завдання двох форм. Завдання 1–4 на заповнення пропусків у мікротекстах (підзавдання 1–10), в яких треба доповнити речення в мікротексті словами / словосполученнями з наведених варіантів. Завдання 5–27 є завданнями з вибором однієї правильної відповіді. Кожне завдання цього типу має основу та чотири варіанти відповіді, з яких лише один правильний. Максимальна кількість балів, яку можна набрати, правильно виконавши всі завдання тесту, — 33 бали: 10 балів за виконання завдань 1–4 (підзавдань 1–10) і 23 бали за виконання завдань 5–27.

Аналіз специфіки побудови блоків завдань ТЗНК й типові помилки при виконанні завдань вступних випробувань до магістратури у 2024 році також представлені в офіційному звіті УЦОЯО [8]. Деякі статистичні характеристики ТЗНК 2024 року наведено в [7], а окремих завдань — у [8]. Водночас, крім аналітики, результати якої представлено в офіційних звітах, розробникам завдань необхідно проводити постійні ґрунтовні дослідження характеристик як тесту загалом, так і кожного завдання зокрема. Для цього використовують методи статистичного аналізу, поєднуючи їх з основними підходами, що застосовують у тестології.

Мета статті: дослідити властивості й обґрунтувати можливості застосування для оцінювання характеристик тесту загальної навчальної компетентності різних математико-статистичних методів. Інтерпретувати отримані результати й розробити рекомендації щодо вибору підходів до оцінювання основних показників якості для окремих тестових завдань і тестів загалом.

Основний матеріал і результати

Для дослідження було обрано дані, що містять результати виконання ТЗНК учасниками однієї із сесій вступних випробувань 2024 року [9], під

час якої вступники виконували завдання тестового варіанта, що оприлюднений в офіційному звіті УЦОЯО [8].

У цій сесії взяло участь 16729 абітурієнтів. Оскільки максимальна кількість балів, яку можна отримати за тест становить 33 й правильна відповідь на кожне із завдань або підзавдань тесту оцінюється в 1 бал, то в подальшому викладі ми будемо вважати окремим завданням кожне завдання або підзавдання тесту, за яке тестований міг отримати 1 або 0 балів.

Результати відповідей вступників на завдання ТЗНК представимо у вигляді матриці $X = \|x_{ij}\|$, $i = \overline{1, n}$, $j = \overline{1, m}$, де $n = 16729$ — обсяг вибірки, $m = 33$ — кількість завдань ТЗНК,

$$x_{ij} = \begin{cases} 0, & \text{якщо } i\text{-ий абітурєнт дав неправильну відповідь на } j\text{-те завдання,} \\ 1, & \text{якщо } i\text{-ий абітурєнт дав правильну відповідь на } j\text{-те завдання.} \end{cases}$$

Величина $y_i = \sum_{j=1}^m x_{ij}$ характеризує сумарний тестовий бал i -го абітурієнта, $0 \leq y_i \leq 33$, $\forall i = \overline{1, n}$.

Нормальність розподілу результатів, перевірка й інтерпретація

Основними найпростішими статистичними характеристиками тесту є:

- середнє вибіркове значення: $\bar{x} = \frac{\sum_{i,j} x_{ij}}{nm} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$;
- стандартне відхилення: $\sigma = \sqrt{\frac{\sum_{i,j} (x_{ij} - \bar{x})^2}{nm}}$;
- асиметрія: $As = \frac{\sum_{i,j} (x_{ij} - \bar{x})^3}{nm\sigma^3}$;
- ексцес: $Ek = \frac{\sum_{i,j} (x_{ij} - \bar{x})^4}{nm\sigma^4} - 3$.

Два останні показники використовують для оцінки відмінності отриманого розподілу результатів тестування від нормального. Для досліджуваної вибірки було отримано такі числові характеристики.

Таблиця 1. Основні числові характеристики розподілу результатів ТЗНК

Обсяг вибірки	Середнє значення	Стандартне відхилення	Асиметрія	Ексцес
16729	16,10	4,83	0,08	2,62

Перевірку нормальності розподілу результатів досить часто виконують за допомогою критерію Колмогорова–Смірнова–Ліллієфорса [10], який є узагальненням критерію Колмогорова для випадку, коли необхідно перевірити складну гіпотезу про узгодженість вибіркових даних з нормальним розподілом за умови, що параметри нормального розподілу невідомі й оцінюються за цією ж вибіркою. Завдяки своїм властивостям, на сьогодні цей критерій є одним з найпоширеніших в практичних застосуваннях.

Використовуючи критерій Колмогорова-Смірнова-Ліллєфорса та програмне середовище SPSS, перевіримо гіпотезу про нормальність розподілу результатів тестування (рис. 1).

One-Sample Kolmogorov-Smirnov Normal Test Summary		
Total N		16729
Most Extreme Differences	Absolute	,047
	Positive	,047
	Negative	-,040
Test Statistic		,047
Asymptotic Sig. (2-sided test)		,000 ^a
a. Lilliefors Corrected		

Рис. 1. Тест Колмогорова-Смірнова-Ліллєфорса для розподілу результатів тестування

Отже, найбільше додатне відхилення емпіричних значень досліджуваного розподілу від теоретичних (нормальних) дорівнює 0,047, а найменшу від’ємне $-0,040$, а отже, спостережуване значення статистики $K_{emp} = \max_i |F_n(y_i) - F(y_i)| = 0,047$, де $F_n(y)$ — емпірична функція розподілу, $F(y)$ — функція нормального розподілу з параметрами, визначеними на основі вибірки (рис. 1). Критичне значення тесту, аналогічно до критичного значення розподілу Колмогорова-Смірнова, обернено пропорційне квадратному кореню з обсягу вибірки. Тому для великих вибірок тест Колмогорова-Смірнова-Ліллєфорса є дуже чутливим до наявності принаймні одного відхилення, яке перевищує критичне значення. У нашому випадку критичне значення приблизно становить $K_n = 0,0068$ і є меншим за емпіричне. Відповідно висновок, який ми отримуємо з використанням автоматизованої перевірки, говорить про те, що нульова гіпотеза про нормальність розподілу відхиляється (рис. 1, останній рядок), оскільки для рівня значущості 0,05 асимптотичне значення ймовірності того, що емпіричне значення відхилиться від теоретичного не більше ніж на K_n не перевищує 0,0005. Графічне представлення розподілу результатів та відповідного теоретичного нормального розподілу наведено на рис. 2.



Рис. 2. Розподіл учасників однієї сесії ЄВІ-2024 за результатами виконання завдань ТЗНК

Для результатів ТЗНК характерним є:

- великий обсяг вибірок, що, як зазначено вище, призводить до зменшення критичних значень тесту Колмогорова–Смірнова–Ліллєфорса;
- наявність окремих точкових викидів здебільшого в групах тестованих, які набрали найбільшу (30–33) або найменшу (0–3) кількість тестових балів.

Унаслідок цього перевірка нормальності за одним з найпоширеніших критеріїв, який до того ж реалізований у статистичних програмних засобах, може призводити до відхилення гіпотези про нормальність, що є підставою вважати тест невалідним [12].

Водночас на практиці можуть виникати принципово різні ситуації — результати можуть очевидно не мати нормального розподілу (рис. 3) або мати розподіл досить близький до нормального (рис. 2). Водночас в обох випадках результати перевірки гіпотези про нормальність будуть аналогічними.

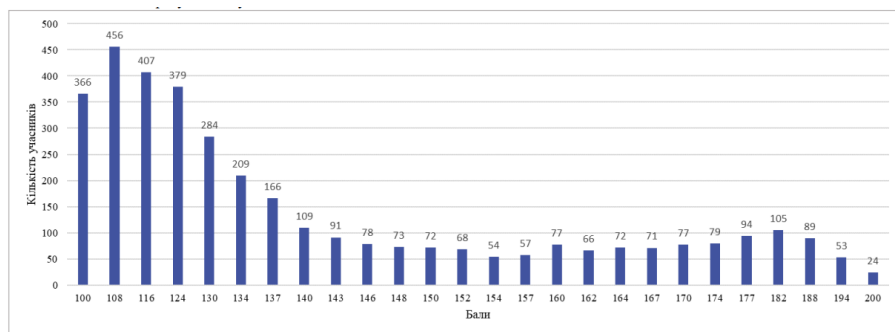


Рис. 3. Розподіл учасників ЄВІ-2024 за результатами виконання завдань з німецької мови [7]

Отже, формальне відхилення гіпотези про нормальність на основі застосування деякого статистичного критерію не може бути єдиним й основним аргументом для визнання його невалідним. Для швидкої візуальної діагностики обґрунтованості припущення нормальності розподілу можна використовувати Q-Q-діаграми (квантиль-квантиль діаграми). Результат побудови такої діаграми для досліджуваної вибірки наведено на рис. 4.

Найсуттєвіші відхилення точок від прямої спостерігаємо для низьких (≤ 6) й високих (≥ 29) результатів, що свідчить про наявність окремих викидів статистичних даних. У цьому випадку критерії, що ґрунтуються на оцінці максимальних відхилень (типу Колмогорова–Смірнова), застосовувати недоцільно.

Існує велика кількість інших критеріїв, які можуть бути використані для перевірки нормальності. на сьогодні доведено [11, 13], що найбільш потужним критерієм для перевірки гіпотези про нормальність розподілу є критерій Шапіро–Вілکا, який використовує відношення оптимальної незміщеної лінійної оцінки дисперсії до її оцінки методом максимальної прав-

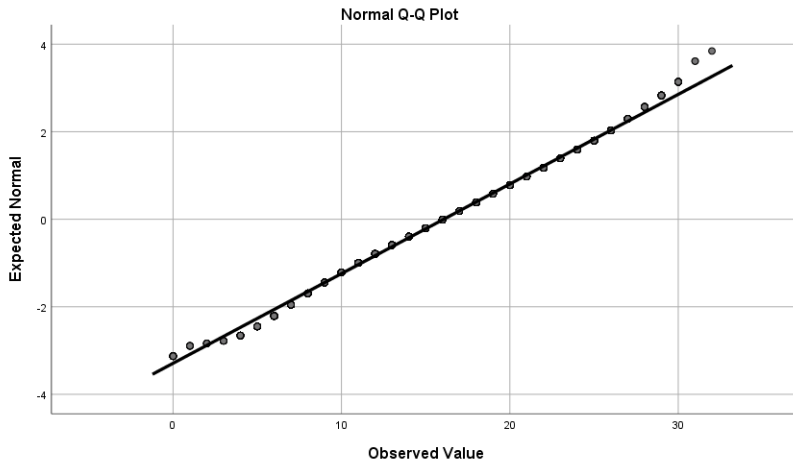


Рис. 4. Q-Q-діаграма за припущення нормальності розподілу результатів тестування

доподібності. Статистика критерію визначається наступним чином:

$$W = \frac{(\sum_{i=1}^n a_i y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad \text{де } a_i = \frac{m^T V^{-1}}{(m^T V^{-1} m)^{1/2}},$$

$m^T = (m_1, \dots, m_n)$ — вектор математичних сподівань незалежних однаково розподілених випадкових величин, що мають нормальний розподіл, V — коваріаційна матриця. Даний критерій має ряд спрощень для великих виборок, зокрема критерій Шапіро-Франція [13].

Tests of Normality			
	Shapiro-Wilk		
	Statistic	df	Sig.
VAR Y	,987	16729	,081

Рис. 5. Тест Шапіро-Вілка для розподілу результатів тестування

Значення статистики критерію Шапіро-Вілка $W_{emp} = 0,987$, його рівень значущості при цьому становить 0,081 (рис. 5). Отже, з ймовірністю не меншою ніж 0,95, ми можемо прийняти нульову гіпотезу про нормальність розподілу результатів тестувань.

Зауважимо, що загальноприйнята ідея про те, що нормальність розподілу результатів є необхідною умовою валідності тесту ґрунтується на евристичному уявленні про універсальність нормального розподілу, що впливає з центральної граничної теореми, а точніше — на використанні оберненого до неї твердження: якщо розподіл випадкової величини, яка є сумою певної кількості випадкових величин (результатів повторних випробувань), виявився близьким до нормального, то результати випробувань були незалежними й однаково розподіленими. Зрозуміло, що такий висновок не є строго обґрунтованим. Тому в практичних дослідженнях можна використовувати різні критерії, аналізувати отримані результати та можливі причини відхилення гіпотези щодо нормальності розподілу.

Числові характеристики розподілу — асиметрія і ексцес також можуть використовуватись для аналізу властивостей тесту. У нашому випадку (табл. 1) асиметрія додатна й близька до 0. Це свідчить про те, що рівень складності близький до оптимального (трохи вищий за оптимальний). Ексцес додатний, що може свідчити про недостатню розподільну здатність тесту. Це впливає з того, що при додатному ексцесі крива емпіричного розподілу має вершину (точку максимуму), ордината якої більша за ординату вершини теоретичного нормального розподілу, відповідно емпірична ймовірність (частота) попадання вибірових значень в околі математичного сподівання (абсциси точки максимуму) більша ніж відповідна теоретична ймовірність. З практичної точки зору це може свідчити про певну «скупченість» вибірових даних в околі середнього значення. Причиною додатного значення ексцеса розподілу результатів тестування може бути те, що тест умовно можна поділити на дві частини — легку й складну або оптимальну й складну.

α -Кронбаха та інші показники надійності тесту

Нехай для i -го тестованого його оцінка за результатами спостереження ξ_i є сумою справжньої оцінки τ_i та похибки вимірювання ε_i , тобто $\xi_i = \tau_i + \varepsilon_i$. Припустимо, що $Cov(\xi_i, \tau_i) = 0$ і $M(\varepsilon_i) = 0$. Тоді для дисперсій випадкових величин $\xi = (\xi_i)_{i=1}^n$, $\tau = (\tau_i)_{i=1}^n$, $\varepsilon = (\varepsilon_i)_{i=1}^n$ виконується рівність: $\sigma_\xi^2 = \sigma_\tau^2 + \sigma_\varepsilon^2$.

Надійність тесту (окремого завдання) r визначається як відношення дисперсії справжньої оцінки до дисперсії спостережуваного значення оцінки: $r = \frac{\sigma_\tau^2}{\sigma_\xi^2}$.

На практиці розглядають кілька підходів для визначення надійності, зокрема: ретестова надійність (передбачає проведення повторних досліджень), надійність паралельних форм (передбачає створення ідентичних за характеристиками паралельних варіантів тестових завдань), надійність частин тесту (передбачає поділ тесту на частини і перевірку узгодженості результатів між частинами і з тестом загалом). Специфіка підготовки й проведення ТЗНК передбачає можливість перевірки надійності тільки шляхом поділу тесту на частини.

З означення величини r випливає, що $r = 1 - \frac{\sigma_\varepsilon^2}{\sigma_\xi^2}$, а отже, $r \in [0, 1]$.

При цьому $r = 1$ тоді і тільки тоді, коли $\sigma_\varepsilon^2 = 0$, тобто $\varepsilon = C = const$. Враховуючи умову $M(\varepsilon) = 0$, маємо $\varepsilon = 0$, тобто справжня оцінка співпадає з вимірюною. Цей випадок відповідає абсолютній надійності тесту й з цього факту випливає оцінка міри надійності через близькість величини r до 1.

Розглянемо які статистичні процедури можна застосувати для оцінки надійності тесту.

При поділі тесту на дві рівні за максимально можливою кількістю балів частини в якості статистичної оцінки ε_i розглядають величину $\hat{\varepsilon}_i = x_{1i} - x_{2i}$, де x_{1i} — оцінка i -го тестованого на першу частину тесту, а x_{2i} —

за другу. Тоді статистичною оцінкою надійності тесту є величина

$$r_2 = 1 - \frac{\sigma_{\varepsilon}^2}{\sigma_{\xi}^2} = 1 - \frac{\sum_{i=1}^n (x_{1i} - x_{2i} - \bar{x}_1 + \bar{x}_2)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (1)$$

Оскільки величина r_2 залежить від відношення дисперсій, то для перевірки статистичної значущості отриманих результатів і визначення меж, у яких може змінюватись r_2 , можна скористатись критерієм Фішера. Для ілюстрації застосування такого підходу розглянемо досліджувану вибірку. Виділимо з отриманих результатів дві рівні за кількістю завдань частини: 15 завдань вербально-комунікативного блоку й 15 завдань логіко-аналітичного (3 завдання вербально-комунікативного блоку не розглядались).

Критичні значення розподілу Фішера для чисел степенів свободи $df_1 = 16728$, $df_2 = 14$ в залежності від рівня значущості наведені в таблиці 2. На основі них можна визначити межі для емпіричного значення показника надійності. Наприклад, якщо емпіричне значення r_2 належить проміжку $[0,53; 0,66)$, то з рівнем значущості не меншим ніж 0,05 (але не більшим ніж 0,01) гіпотеза про рівність дисперсій σ_{ε}^2 і σ_{ξ}^2 відхиляється, тобто теоретичне значення надійності суттєво відрізняється від 0.

Таблиця 2. Критичні значення розподілу Фішера та межі для показника надійності

Рівень значущості	0,5	0,10	0,05	0,01	< 0,01
Критичне значення розподілу Фішера	0,95	0,56	0,47	0,33	> 0,33
Межі, в яких змінюється показник надійності	[0; 0,15)	[0,15; 0,44)	[0,44; 0,53)	[0,53; 0,66)	[0,66; 1]
Висновки щодо рівня надійності	Неприйнятний	Низький	Середній	Достатній	Високий

Для досліджуваної вибірки емпіричне значення показника надійності r_2 становить 0,712, що свідчить про високий рівень надійності тесту.

Зрозуміло, що описаний підхід можна застосовувати для будь-яких частин тесту. Загалом існує C_{2k}^k способів розподілу $2k$ питань тесту на дві частини. Розглянувши усі, можна визначити мінімальне й максимальне значення надійності:

$$r_{2,min} = \min_{j=1, C_{2k}^k} \{r_{2,j}\}, \quad r_{2,max} = \max_{j=1, C_{2k}^k} \{r_{2,j}\}.$$

Перевірка надійності тесту шляхом поділу його на частини може також здійснюватися за допомогою оцінки кореляції між отриманими складовими. У випадку наявності суттєвого прямого зв'язку тест вважається надійним.

Нехай X_j — спостережуване значення оцінок для j -го питання тесту, $j = \overline{1, m}$, $X = X_1 + \dots + X_m$. Позначимо $\sigma_{ij} = \text{cov}(X_i, X_j)$, а якщо $i = j$, то $\sigma_{ii} = \sigma_i^2 = D(X_i)$. Тоді

$$\sigma_X^2 = \sum_{i=1}^m \sigma_i^2 + \sum_{i \neq j} \sigma_{ij}.$$

Середня міжгрупова коваріація:

$$\bar{\sigma}_{ij} = \frac{\sum_{i \neq j} \sigma_{ij}}{m(m-1)}.$$

Тоді

$$1 - \frac{\sum_{i=1}^m \sigma_i^2}{\sigma_X^2} = \frac{\sum_{i \neq j} \sigma_{ij}}{\sigma_X^2} \Rightarrow 1 - \frac{\sum_{i=1}^m \sigma_i^2}{\sigma_X^2} = \frac{m(m-1) \bar{\sigma}_{ij}}{\sigma_X^2}.$$

Ліва частина останньої рівності є аналогом формули (1), де замість дисперсії різниць між балами за різні частини тесту використана сума дисперсій результатів за кожним компонентом. Очевидно, що

$$1 - \frac{\sum_{i=1}^m \sigma_i^2}{\sigma_X^2} < \frac{m^2 \bar{\sigma}_{ij}}{\sigma_X^2} \quad \text{та} \quad \frac{m^2 \bar{\sigma}_{ij}}{\sigma_X^2} = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^m \sigma_i^2}{\sigma_X^2}\right).$$

Величину

$$\rho_\alpha = \frac{m^2 \bar{\sigma}_{ij}}{\sigma_X^2} = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^m \sigma_i^2}{\sigma_X^2}\right) \quad (2)$$

називають α Кронбаха [14].

Покажемо, що за деяких додаткових припущень α Кронбаха є точною оцінкою показника надійності.

Ми розглядаємо оцінку за результатами спостереження ξ_i як суму справжньої оцінки τ_i та похибки вимірювання ε_i , тобто $\xi_i = \tau_i + \varepsilon_i$.

Припустимо, що

- 1) $\text{Cov}(\tau_i, \varepsilon_i) = 0$ і $\text{Cov}(\tau_i, \varepsilon_j) = 0$;
- 2) $M(\varepsilon_i) = 0$;
- 3) $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$.
- 4) $\tau_i = \mu_i + c$, де μ_i — справжній бал, однаковий для всіх питань тесту, c — деяка константа. Цю умову називають умовою «тау-еквівалентності».

За цих припущень маємо:

$$\begin{aligned} \sigma_i^2 &= D(\mu_i + c + \varepsilon_i) = D(\mu_i) + D(c) + D(\varepsilon_i) \\ &\quad + 2(\text{Cov}(\mu_i, c) + \text{Cov}(\mu_i, \varepsilon_i) + \text{Cov}(c, \varepsilon_i)), \end{aligned}$$

$\text{Cov}(\mu_i, c) = \text{Cov}(c, \varepsilon_i) = \text{Cov}(\mu_i, \varepsilon_i) = D(c) = 0$ в силу припущення 4, отже $\sigma_i^2 = \sigma_{\mu_i}^2 + \sigma_{\varepsilon_i}^2$.

$$\sigma_{ij} = \text{Cov}(\mu_i + c + \varepsilon_i, \mu_j + c + \varepsilon_j) = \text{Cov}(\tau_i + \varepsilon_i, \tau_j + \varepsilon_j) = \sigma_{\mu_i}^2$$

в силу припущень 1 і 3. Тоді $\bar{\sigma}_{ij} = \sigma_{\mu_i}^2$, а $\sigma_\tau^2 = D(\sum_{i=1}^m \mu_i) = m^2 \bar{\sigma}_{ij}$.

Отже, при виконанні наведених вище припущень величина $\rho_\alpha = \frac{m^2 \bar{\sigma}_{ij}}{\sigma_X^2}$ дає точну оцінку значення надійності тесту.

На практиці ці припущення часто не виконуються, а їх перевірка вимагає ґрунтовних додаткових досліджень. Особливо обмежувальним є припущення тау-еквівалентності, оскільки воно вимагає по суті певної однорідності тестових балів, якої рідко досягають на практиці. З наведених вище міркувань випливає, що за умови порушення припущення тау-еквівалентності величина α Кронбаха буде меншою за істинне значення надійності.

Крім α Кронбаха існує цілий ряд аналогічних показників з подібними властивостями, зокрема:

- Показник Кудера–Річардсона:

$$\rho_{KR} = \frac{m}{m-1} \cdot \left(1 - \frac{\sum_{i=1}^m p_i (1-p_i)}{\sigma_X^2} \right),$$

де p_i — частка правильних відповідей на i -те питання, тобто показник складності питання тесту. Очевидно, що показник Кудера–Річардсона співпадає з α Кронбаха у випадку біноміального розподілу результатів.

- Лямбди Гутмана. Гутман запропонував і дослідив шість показників надійності: $\lambda_1, \lambda_2, \dots, \lambda_6$. При цьому $\lambda_3 = \rho_\alpha$. на сьогодні найчастіше використовується

$$\lambda_4 = \frac{4Cov(X_1, X_2)}{\sigma_X^2},$$

де X_1, X_2 — бали за першу і другу частини розщепленого тесту.

Проведемо аналіз надійності тесту ТЗНК на основі даних досліджуваної вибірки.

Reliability Statistics				
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items	Lambda	
			1	,708
			2	,736
			3	,730
			4	,679
			5	,722
			6	.
			N of Items	33
,730	,730	33		

Рис. 6. Результати обчислення α Кронбаха та λ Гутмана

Традиційно вважають, що при значенні $\rho_\alpha \geq 0,75$ тест є надійним. Але оскільки значення ρ_α оцінюють статистично, то очевидно, що існує суттєва залежність від обсягу вибірки й кількості завдань у тесті. А отже, відповідні критичні значення мають це враховувати. Використаємо запропонований вище підхід для визначення меж загального показника надійності для уточнення меж, у яких може змінюватися α Кронбаха в нашому дослідженні.

Таблиця 3. Межі для α Кронбаха

Рівень значущості	0,5	0,10	0,05	0,01	<0,01
Межі, в яких змінюється показник надійності	[0; 0,05)	[0,05; 0,45)	[0,45; 0,54)	[0,54; 0,69)	[0,69; 1]
Висновки щодо рівня надійності	Неприйнятний	Низький	Середній	Достатній	Високий

Отже, отримане емпіричне значення α Кронбаха для досліджуваної вибірки відповідає високому рівню надійності тесту.

Крім загальної характеристики тесту α Кронбаха може бути застосована для аналізу окремих завдань. Для цього визначають α Кронбаха тесту із вилученням результатів, отриманих для одного із завдань. Якщо при цьому α Кронбаха суттєво зростає, то дане завдання знижує надійність тесту й доцільно розглянути питання про його вилучення.

Результати такої перевірки наведено в таблиці 4. Для ілюстрації ми залишили тільки частину завдань, вилучивши ті, що мали гарні статистичні характеристики.

На основі результатів проведеного аналізу можна зробити висновки, що підвищення α Кронбаха відбувається при вилученні завдань 1, 6 і 30, а найменш суттєве зниження — при вилученні завдань 7, 16 і 27. Крім цього результати виконання завдань 6 і 30 мають від'ємний показник кореляції із загальним тестовим балом, а завдання 1, 7, 16 і 27 — найнижчі додатні показники кореляції.

Зауважимо, що розрахунок показника кореляції результатів виконання окремих завдань для випадку вилучених елементів відрізняється від розрахунку інших показників кореляції — Пірсона [8], Спірмена (рис. 7) або Кендела. Ці коефіцієнти підтверджують наявність суттєвого прямого зв'язку для всіх завдань, крім 30 (див. результати обчислень коефіцієнта Спірмена, рис. 7, для коефіцієнтів Пірсона й Кендела результати аналогічні).

Також доцільно аналізувати зміну α Кронбаха за умови вилучення блоків завдань. У нашому випадку ми розглянули чотири блоки й отримали такі результати:

- за умови виключення 10 завдань на відновлення цілісності мікротекстів: $\rho_\alpha = 0,667$;
- за умови виключення 8 завдань на аналіз двох текстів: $\rho_\alpha = 0,658$;
- за умови виключення 6 логічних завдань: $\rho_\alpha = 0,684$;
- за умови виключення 9 логічних завдань: $\rho_\alpha = 0,689$.

Також були розраховані показники зміни α Кронбаха для окремих змістових блоків тесту (рис. 8).

Отримані результати свідчать про те, що вилучення жодного з блоків завдань не підвищує загальну надійність тесту, а показники надійності

Таблиця 4. Результати аналізу α Кронбаха методом вилучення окремих завдань

Item	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
VAR00001	15,4928	23,117	,100	,732
VAR00002	15,6648	22,116	,321	,719
VAR00003	15,3779	22,431	,273	,722
VAR00004	15,4736	22,093	,326	,718
VAR00005	15,3941	22,774	,189	,727
VAR00006	15,9449	23,817	-,027	,735
VAR00007	15,5664	22,750	,176	,728
...				
VAR00014	15,3463	21,969	,397	,715
VAR00015	15,5759	22,000	,339	,718
VAR00016	15,9379	23,257	,144	,728
VAR00017	15,2922	22,549	,282	,722
VAR00018	15,5829	22,633	,201	,726
VAR00019	15,5082	22,285	,279	,721
...				
VAR00024	15,5145	22,512	,229	,725
VAR00025	15,3946	21,993	,370	,716
VAR00026	15,8184	22,763	,215	,725
VAR00027	15,6935	22,888	,155	,729
VAR00028	15,8469	22,523	,290	,721
VAR00029	15,5763	21,988	,341	,717
VAR00030	15,7753	24,288	-,146	,745

окремих блоків є достатніми або середніми. Найвищим є показник α Кронбаха для блоку завдань на аналіз двох текстів, а найнижчим — для блоку логічних завдань. Але враховуючи, що значення α Кронбаха може помітно змінюватись навіть із вилученням одного завдання, робити однозначні висновки на основі порівняння показників надійності цих блоків не можна. З упевненістю можна стверджувати тільки те, що надійність блоку завдань ТЗНК на аналіз двох текстів (виміряна за допомогою α Кронбаха) вища, ніж надійність блоку завдань на відновлення цілісності мікротекстів та аналітичного блоку.

Аналіз деяких характеристик окремих завдань тесту

У цій частині роботи проаналізуємо зміст і психометричні характеристики завдань, які виявились такими, що знижують загальну надійність тесту і результати їх виконання мають поганий кореляційний зв'язок з результатами виконання тесту загалом. Наведемо й проаналізуємо такі характеристики завдань:

Spearman's rho												
	1	2	3	4	5	6	7	8	9	10	11	12
Correlation Coefficient	,194**	,409**	,356**	,415**	,275**	,035**	,266**	,277**	,394**	,400**	,398**	,291**
Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	13	24	15	16	17	18	19	20	21	22	23	24
Correlation Coefficient	,322**	,474**	,428**	,204**	,354**	,300**	,370**	,320**	,316**	,289**	,376**	,318**
Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	25	26	27	28	29	30	31	32	33			
Correlation Coefficient	,454**	,285**	,239**	,354**	,432**	-,064**	,302**	,323**	,303**			
Sig. (2-tailed)	,000	,000	,000	,000	,000	,000	,000	,000	,000			

** Correlation is significant at the 0.01 level (2-tailed).

Рис. 7. Результати обчислення і перевірки значущості коефіцієнта кореляції Спірмена між результатами виконання окремих завдань тесту й загальним балом

- складність тестового завдання (*P-value*) — відношення (у відсотках) кількості балів, набраних усіма учасниками за виконання цього завдання, до максимальної кількості балів, яку вони могли б отримати за його виконання [8];
- дискримінативність (розподільна здатність) тестового завдання (*D-index*) — різниця складності завдання для сильної та слабкої (добре й погано підготовленої) груп учасників тестування [8];
- коефіцієнт кореляції Пірсона;
- діаграми розподілу частот вибору кожного з варіантів відповідей на завдання для різних груп тестованих. Для цього аналізу всі тестовані були розподілені на сім груп: група 1 — ті, хто набрав за результатами тестування від 0 до 3 балів, група 2 — від 4 до 8, група 3 — від 9 до 13, група 4 — від 14 до 18, група 5 — від 19 до 23, група 6 — від 24 до 28, група 7 — від 29 до 33 балів.

Зазначимо, що ми використовуємо наскрізну нумерацію завдань тесту (від 1 до 33), тоді як в наведеному варіанті [8] деякі завдання об'єднані в групи й пронумеровані від 1 до 27. Відповідність нумерації завдань можна встановити зокрема за таблицею 5.1.3.1 [8].

У першому блоці вербально-комунікативного компонента ТЗНК, завдання якого передбачають виявлення особою, що тестується, здатності точно й ефективно подавати інформацію в текстових повідомленнях, проблемними виявилися три завдання — 1, 6 та 7.

Reliability Statistics							
Cronbach's Alpha	N of Items	Cronbach's Alpha	N of Items	Cronbach's Alpha	N of Items	Cronbach's Alpha	N of Items
,458	10	,503	8	,355	6	,454	9

Рис. 8. Результати розрахунку α Кронбаха для окремих блоків завдань тесту

Завдання 1 перевіряло вміння розрізняти значення близьких за звуковим складом і написанням, але різних за значенням слів і вживати їх у відповідному контексті. Як свідчать представлені в звіті дані, дистрактор Г іспитовані майже не вибирали (0,6%), а дистрактор Б вибрали всього 8,8% учасників тестування. Натомість основний вибір іспитовані робили між варіантом А (55,7%) та варіантом В (32,8%). Відмінність між запропонованими в зазначених варіантах іменниках полягає в тому, що вони утворені від різних форм дієслова, а саме: недоконаного та доконаного виду. Водночас саме дієслова недоконаного виду виражають незавершену дію без указівки на її результат, тобто ключ від дистрактора вирізняє на семантичному рівні ознака *‘процес’*, яку змогли визначити не всі іспитовані. Це завдання мало оптимальний рівень складності, середній рівень дискримінативності, додатний показник кореляції. Загалом розподіл частот вибору варіантів відповідей (рис. 9) свідчить про те, що крива, що відповідає ключу (варіант А) є висхідною, тоді як інші криві, які відповідають дистракторам, спадними для всіх груп тестованих, що набрали не менше 4 балів. Це свідчить про належні психометричні характеристики завдання.

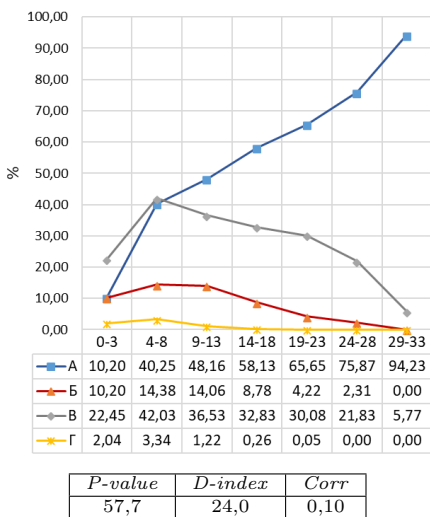


Рис. 9. Психометричні характеристики завдання 1

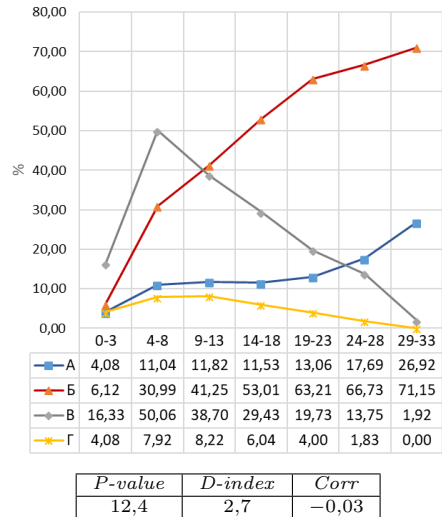


Рис. 10. Психометричні характеристики завдання 6

Завдання 6, яке перевіряло вміння використовувати розділові знаки відповідно до норм української мови, виявилось складним. З ним успішно впоралося лише 12% тестованих. Причин може бути декілька: по-перше, недостатній рівень володіння пунктуаційними нормами української мови; по-друге, нетиповість завдання, адже обрати необхідно було варіант без розділового знака. Імовірно, ускладнення спричинила також невідповідність до такого типу завдань, адже завдання на пунктограми в ТЗНК було використано вперше. Це завдання мало дуже високий рівень складності та, як наслідок, низьку дискримінативність, близький до 0 від'ємний показник кореляції. Аналіз частотної діаграми показує (рис. 10), що поведінку, характерну для ключа, демонструють дві криві — ті, що відповідають варіантам відповідей А та Б. Це може свідчити про те, що один з дистракторів

виявився привабливішим за ключ для усіх категорій тестованих. Валідувати таке завдання можна зміною одного з дистракторів.

Завдання 7 перевіряло знання питомої лексики й уміння використовувати її в контексті. Складність завдання обумовлена тим, що опорні слова, за якими можна визначити правильну відповідь, розташовані за пропуском у мікротексті, а не перед ним. Імовірним є також те, що значна частина учасників тестування обирала варіант В (43,3%) через наявність усталеного (клішованого) асоціативного зв'язку між словами *потреба* і *споживач* у свідомості носіїв мови, та водночас відсутність такого зв'язку між словами *рішення* і *споживач*. Це завдання мало оптимальний рівень складності, хороший рівень дискримінативності, додатний показник кореляції. Розподіл частот вибору варіантів відповідей (рис. 11) свідчить про те, що крива, що відповідає ключу (варіанта Б) є висхідною, також можемо спостерігати високу привабливість дистрактора В для груп тестованих, які набрали менше ніж 18 балів. Загалом завдання має належні характеристики, його валідизація може полягати в підборі більш однорідних дистракторів.

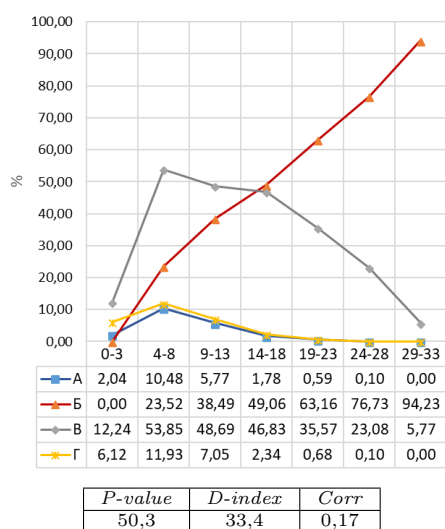


Рис. 11. Психометричні характеристики завдання 7

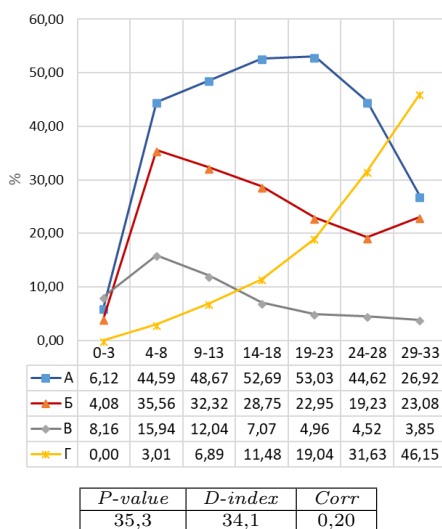


Рис. 12. Психометричні характеристики завдання 16

Як свідчать результати виконання завдань другого блоку вербально-комунікативного компонента ТЗНК, які передбачали виявлення тестованим спроможності аналізувати, інтерпретувати й оцінювати текстову інформацію, найбільші труднощі викликало завдання, що стосувалося використання авторами запропонованих текстів мовних прийомів і засобів виразності. Так, один із текстів містив іронію та сарказм, проте більшості учасників тестування не змогла визначити оцінне, глузливо-критичне ставлення, до зображуваного. Це завдання мало високий рівень складності, хороший рівень дискримінативності, додатний статистично значущий показник кореляції. Жодна з цих характеристик не вказує на наявність проблем, хоча, як зазначалось вище, аналіз α Кронбаха методом виключення елементів показав можливе підвищення надійності тесту загалом при виключенні цього завдання. Аналіз діаграми розподілу частот вибору варіан-

тів відповідей (рис. 12) демонструє те, що дистрактор А обирали найчастіше усі групи тестованих, крім груп 1 (найгірші результати) і 7 (найкращі результати), також нетиповим є зростання частоти вибору дистрактора Б для групи 7. При цьому варіант Г демонструє типове для ключа зростання частотної кривої. Отже, це завдання, не зважаючи на непогані статистичні характеристики, потребує вдосконалення, зокрема зміни дистракторів А і Б.

Завдання 27 другого блоку логіко-аналітичного компонента ТЗНК стосувалося ситуації 1, у якій перевірялось уміння тестованих визначати структуру скінченних множин, які задовольняють заданим умовам, перевіряти виконання заданої системи умов, встановлювати істинність або хибність тверджень з урахуванням додаткової інформації. Аналіз відповідей на завдання 27 дає підстави констатувати, що в абітурієнтів виникли складнощі з перевіркою істинності тверджень, сформульованих у формі «якщо ..., то ...», які передбачають врахування кількох умов. Аналіз відповідей показав, що при встановленні істинності двох наведених тверджень типовою помилкою було визначення твердження ІІ як хибного, тоді як з встановленням хибності твердження І у більшості тестованих проблем не виникло. на нашу думку це пов'язано з тим, що для встановлення істинності твердження ІІ необхідно було залучити аналітичні комбінаторні міркування, які викликали певні складнощі у тестованих. Це завдання мало високий рівень складності, середній рівень дискримінативності, додатний показник кореляції. Розподіл частот вибору варіантів відповідей (рис. 13) свідчить про те, що крива, що відповідає ключу (варіанта Б) є висхідною, також можемо спостерігати немонотонність кривої, що відповідає дистрактору Г, причому вона є висхідною для груп тестованих, що набрали від 4 до 23 балів. Загалом це завдання має належні характеристики, його валідизація може полягати у зміні твердження ІІ в стимулі завдання.

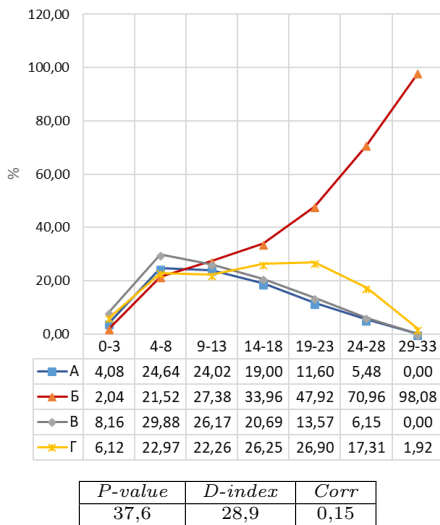


Рис. 13. Психометричні характеристики завдання 27

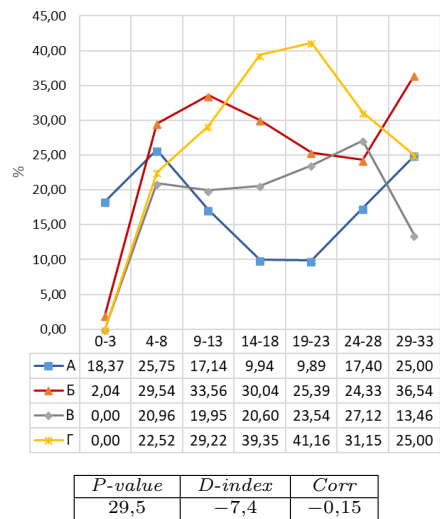


Рис. 14. Психометричні характеристики завдання 30

Завдання 30 стосувалося другої ситуації і перевіряло вміння тестованих розв'язувати задачі комбінаторного змісту, виконувати оцінку числових значень невідомих величин. Аналіз відповідей на завдання 30 (рис. 14) дає підстави констатувати, що це завдання виявилось складним для більшості абітурієнтів, частотна діаграма демонструє ознаки випадкового вибору варіантів, що свідчить про те, що більшість тестованих «вгадували» правильну відповідь або обирали її навмання. Однією з причин цього може бути неможливість отримати правильну відповідь шляхом евристичних міркувань, необхідність застосовувати математичні розрахунки, використовуючи найпростіші факти комбінаторики (теореми множення і додавання), до чого більшість тестованих не були готовими.

Висновки

Аналіз характеристик тестів, які використовують для масових зовнішніх незалежних оцінювань, є складним і дуже важливим завданням. на сьогодні не існує усталених підходів використання математичних методів для його вирішення. Тому на практиці розробники використовують певні конвенційно прийняті методики, які через низку різних причин часто застосовують формально. Розуміння теоретичних основ, меж застосовності та припущень, які забезпечують адекватність результатів, отриманих за допомогою математико-статистичних підходів є необхідною умовою ефективності їх застосування. Проведене дослідження дає змогу зробити висновки й сформулювати рекомендації:

- 1) Висновки щодо перевірки нормальності розподілу результатів тестувань доцільно інтерпретувати таким чином: якщо нормальність підтверджено з достатнім рівнем надійності, це свідчить про непогані статистичні характеристики тесту й можливість застосовувати в подальших дослідженнях параметричні критерії, а якщо ні, то варто проаналізувати причини відхилення гіпотези, серед яких основними можуть бути: суттєве відхилення емпіричних даних від нормального розподілу, наявність окремих нечисленних викидів, завищений рівень надійності для якого визначалось критичне значення тощо. В усіх випадках, крім явної невідповідності емпіричного розподілу нормальному, доцільно провести додаткові дослідження: застосувати різні критерії, проаналізувати Q-Q- та P-P-діаграми, перевірити гіпотезу про наявність лог-нормального або іншого аналогічного типу розподілу.
- 2) Враховуючи специфіку результатів ТЗНК, можна рекомендувати використовувати для перевірки нормальності тести Шапіро-Вілکا або Андерсона-Дарлінга, які краще враховують наявність випадкових відхилень емпіричних значень від теоретичних ніж тест Колмогорова-Смірнова-Ліллієфорса.
- 3) α Кронбаха є одним з показників надійності тесту, але його емпіричне значення є незміщеною оцінкою справжньої надійності тільки за виконання низки припущень, про які йшлося в статті. В інших випадках α Кронбаха набуває значень менших за істинну надійність. Крім α Кронбаха можна використовувати й інші показники надійності тесту. У випадку, коли надійність перевіряється шляхом поділу тести на ча-

стини, можна використовувати показники λ Гутмана, оскільки вони дають можливість знайти верхню і нижню оцінку надійності. Тоді як застосування показника Кудера-Річардсона для аналізу тестів ТЗНК не є виправданим, оскільки використовує додаткове припущення про біноміальність розподілу результатів відповідей на кожне з питань.

- 4) Крім визначення загального показника α Кронбаха, ефективним методом дослідження якості окремих завдань може бути метод визначення α Кронбаха з вилученням елементів, а також обчислення цього показника надійності для окремих змістових блоків тесту.
- 5) Для уточнення результатів аналізу якості окремих завдань тесту ефективним механізмом може бути використання діаграм розподілу частот, з якими обирали кожен з варіантів відповідей різні групи тестованих, впорядкованих за зростанням загального результату виконання тесту.

Список використаних джерел

- [1] Про вищу освіту: Закон України від 1 липня 2014 р. № 1556-VII, URL: <http://zakon3.rada.gov.ua/laws/show/1556-18> (дата звернення: 12.09.2024).
- [2] Методичні рекомендації щодо розроблення стандартів вищої освіти. Наказ Міністерства освіти і науки України від 01.06.2017 р. № 600 (у редакції наказу Міністерства освіти і науки України від 30.04.2020 р. № 584) URL: <https://mon.gov.ua/static-objects/mon/sites/1/vishcha-osvita/rekomendatsii-1648.pdf> (дата звернення: 12.09.2024).
- [3] Концепція зовнішнього незалежного оцінювання із загальної навчальної правничої компетентності для проведення у 2016 році як експерименту вступних випробувань під час вступу на основі ступеня бакалавра на навчання для здобуття ступеня магістра за спеціальністю 081 «Право». 2016. *Вісник «Тестування і моніторинг в освіті»*. № 4–5. С. 7–20.
- [4] LSAT — вступний тест до магістерських програм правничих шкіл (досвід США). 2015 / перекл. з англ. С. А. Раков, В. М. Терещенко, Т. С. Вакулєнко. *Вісник «Тестування і моніторинг в освіті»*. № 11–12. С. 28–43.
- [5] Програма Тесту загальної навчальної компетентності (Наказ МОН України № 158 від 11 лютого 2022 року) URL: <https://testportal.gov.ua/yedynij-vstupnyj-ispit-2/> (дата звернення: 12.09.2024).
- [6] Структура Тесту загальної навчальної компетентності, URL: <https://testportal.gov.ua/wp-content/uploads/2024/03/nakaz-vid-25.03.2024-38.pdf> (дата звернення: 12.09.2024).
- [7] Офіційний звіт УЦОЯО про проведення вступних випробувань до магістратури й аспірантури у 2024 році: єдиного вступного іспиту (ЄВІ) та єдиного фахового вступного випробування (ЄФВВ). Том 1. URL: https://testportal.gov.ua/wp-content/uploads/2024/12/Zvit_YEVI_YEFVV_2024_1.pdf (дата звернення: 12.09.2024).
- [8] Офіційний звіт УЦОЯО про проведення вступних випробувань до магістратури й аспірантури у 2024 році: єдиного вступного іспиту (ЄВІ) та єдиного фахового вступного випробування (ЄФВВ). Том 2. URL:

- https://testportal.gov.ua/wp-content/uploads/2024/12/ZVIT-EVI_2024-Tom_2.pdf (дата звернення: 12.09.2024).
- [9] Деперсоніфікована база даних ЄБІ/ЄФББ URL: <https://drive.google.com/drive/folders/1LCkuTZIFSVHCslZbj5FEKNAY010Z0GWH> (дата звернення: 12.09.2024).
 - [10] Lilliefors H. 1969. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association*, Vol. 64, No. 325. (Mar., 1969), P. 387–389.
 - [11] Razali N. M., Yap Bee Wah. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*. Vol. 2, No. 1, P. 21–33.
 - [12] Булах І.Є., Мруга М.Р. 2006. Створюємо якісний тест : навч. посіб. Київ : Майстер; клас. 160 с.
 - [13] Nornadiiah Mohd Razali1, Yap Bee Wah. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, Vol. 2, No. 1, P. 21–33.
 - [14] Cho E., Kim S. 2015. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2), P. 207–230.

References

- [1] Pro vyshchu osvitu: Zakon Ukrainy vid 1 lypnia 2014 r. № 1556–VII, URL: <http://zakon3.rada.gov.ua/laws/show/1556-18> (data zvernennia: 12.09.2024).
- [2] Metodychni rekomendatsii shchodo rozroblennia standartiv vyshchoi osvity. Nakaz Ministerstva osvity i nauky Ukrainy vid 01.06.2017 r. № 600 (u redaktsii nakazu Ministerstva osvity i nauky Ukrainy vid 30.04.2020 r. № 584) URL: <https://mon.gov.ua/static-objects/mon/sites/1/vishcha-osvita/rekomendatsii-1648.pdf> (data zvernennia: 12.09.2024).
- [3] Kontseptsiiia zovnishnoho nezalezhnogo otsiniuvannia iz zahalnoi navchalnoi pravnychoi kompetentnosti dlia provedennia u 2016 rotsi yak eksperymentu vstupnykh vyprobuvan pid chas vstupu na osnovi stupenia bakalavra na navchannia dlia zdobuttia stupenia mahistra za spetsialnistiu 081 «Pravo». 2016. *Visnyk «Testuvannia i monitorynh v osviti»*. № 4–5. S. 7–20.
- [4] LSAT — vstupnyi test do mahisterskykh prohram pravnychkykh shkil (dosvid SSHA). 2015. / translated from English by S. A. Rakov, V. M. Tereshchenko, T. S. Vakulenko. *Visnyk «Testuvannia i monitorynh v osviti»*. № 11–12. S. 28–43.
- [5] Prohrama Testu zahalnoi navchalnoi kompetentnosti (Nakaz MON Ukrainy № 158 vid 11 liutoho 2022 roku) URL: <https://testportal.gov.ua/yedynyj-vstupnyj-ispyt-2/> (data zvernennia: 12.09.2024).
- [6] Struktura Testu zahalnoi navchalnoi kompetentnosti, URL: <https://testportal.gov.ua/wp-content/uploads/2024/03/nakaz-vid-25.03.2024-38.pdf> (data zvernennia: 12.09.2024).
- [7] Ofitsiyni zvit UTsOIAo pro provedennia vstupnykh vyprobuvan do mahistratury y aspirantury u 2024 rotsi: yedynoho vstupnogo ispytu (IeVI) ta yedynoho fakhovoho vstupnogo vyprobuвання (IeFVV). Vol. 1. URL: https://testportal.gov.ua/wp-content/uploads/2024/12/Zvit_YEVI_YEFVV_2024_1.pdf (data zvernennia: 12.09.2024).

- [8] Ofitsiinyi zvit UTsOIaO pro provedennia vstupnykh vyprobuvan do mahistratury y aspirantury u 2024 rotsi: yedynoho vstupnoho ispytu (IeVI) ta yedynoho fakhovoho vstupnoho vyprobuvannia (IeFVV). Vol. 2. URL: https://testportal.gov.ua/wp-content/uploads/2024/12/ZVIT-EVI_2024-Tom_2.pdf (data zvernennia: 12.09.2024).
- [9] Depersonifikovana baza danykh YeVI/YeFVV URL: <https://drive.google.com/drive/folders/1LCkuTZIFSVHCsIZbj5FEKNAY010Z0GWH> (data zvernennia: 12.09.2024).
- [10] Lilliefors H. 1969. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *Journal of the American Statistical Association*, Vol. 64, No. 325. (Mar., 1969), pp. 387–389.
- [11] Razali N. M., Yap Bee Wah. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*. Vol. 2, No. 1, P. 21–33.
- [12] Bulakh I. Ye., Mruha M. R. 2006. Stvoriuiemo yakisnyi test : navch. posib. Kyiv : Maister;klas. 160 p.
- [13] Nornadiah Mohd Razali1, Yap Bee Wah. 2011. Power comparisons of Shapiro–Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, Vol. 2, No. 1, P. 21–33.
- [14] Cho E., Kim S. 2015. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18(2), P. 207–230.