НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Фізико-математичний факультет

Кафедра математичного аналізу та теорії ймовірностей

«На правах рукопису»
УДК 519.2

До захисту допущено:
Завідувач кафедри
_____ Олег КЛЕСОВ
«___»_____20__ р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-професійною програмою «Страхова та фінансова математика»

зі спеціальності 111 «Математика»

на тему: «Розробка та статистичний аналіз якості покрокових комп'ютерних тестів у системі STACK з математичних дисциплін»

Виконала:
студентка ІІ курсу магістратури, групи ОМ-41мп
Микитенко Вікторія Сергіївна _____

Науковий керівник:
кандидат фізико-математичних наук, доцент
Орловський Ігор Володимирович _____

Рецензент:
кандидат фізико-математичних наук, доцент
Нестеренко Олексій Никифорович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студентка _____

Київ – 2025 року

Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»
Фізико-математичний факультет
Кафедра математичного аналізу та теорії ймовірностей

Рівень вищої освіти – другий (магістерський)

Спеціальність – 111 «Математика»

Освітньо-професійна програма «Страхова та фінансова математика»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олег КЛЕСОВ
«___»_____20__ р.

ЗАВДАННЯ
на магістерську дисертацію студенту
Микитенко Вікторії Сергіївні

1. Тема дисертації «Розробка та статистичний аналіз якості покрокових комп'ютерних тестів у системі STACK з математичних дисциплін», науковий керівник дисертації Орловський Ігор Володимирович, кандидат фізико-математичних наук, доцент, затверджені наказом по університету від «06» листопада 2025 р. №4843-с

2. Термін подання студентом дисертації: 18.12.2025

3. Об'єкт дослідження: покрокові комп'ютерні тести з математичних дисциплін.

4. Предмет дослідження: психометричні характеристики якості покрокових тестових завдань з математичних дисциплін, що оцінюються за допомогою методів класичної теорії тестування (CTT) та моделей сучасної теорії тестування (IRT).

5. Перелік завдань, які потрібно розробити:

1) Ознайомитися з літературою за темою магістерської дисертації. Проаналізувати наукові джерела, присвячені класичній та сучасній теорії тестування, а також сучасним підходам до оцінювання якості комп'ютерних тестів у математичній освіті.
2) Дослідити структуру та логіку побудови покрокових тестових завдань з математичних дисциплін, що використовуються у системі Moodle, зокрема із застосуванням модуля STACK.

3) Здійснити ручний збір результатів виконання покрокових тестових завдань студентами, сформувати вибірку емпіричних даних та виконати їхню попередню підготовку до статистичного аналізу.
4) Провести статистичний аналіз якості покрокових тестових завдань із використанням методів класичної теорії тестів (CTT) та моделей сучасної теорії тестування (IRT).
5) Виконати порівняльний аналіз психометричних характеристик стандартних покрокових тестів і тестів на основі модуля STACK та сформулювати рекомендації щодо їхнього удосконалення.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу: 24 слайди презентації.

7. Орієнтовний перелік публікацій: Микитенко В.С., Орловський І.В. Статистичний аналіз якості покрокових комп'ютерних тестів у системі STACK з математичних дисциплін // XX Міжнародна наукова конференція імені академіка Михайла Кравчука, 17–20 листопада 2025 р. : матеріали конф. – Київ : НТУУ «КПІ ім. І. Сікорського», 2025. – С. 200–201

8. Дата видачі завдання: 03.09.2025

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Постановка, формулювання та обговорення головної задачі магістерської дисертації.	03.09.2025 – 14.09.2025	Виконано
2	Аналіз літератури та публікацій. Вивчення особливостей побудови покрокових тестів у середовищі Moodle. Ознайомлення з надбудовою STACK	15.09.2025 – 28.09.2025	Виконано
3	Аналіз існуючих результатів в класичній (CTT) та сучасній теорії тестів (IRT).	29.09.2025 – 12.10.2025	Виконано
4	Збір результатів тестування, попередня обробка даних з використанням MS Excel.	13.10.2025 – 26.10.2025	Виконано
5	Розробка в R програмного коду для аналізу стандартного покрокового тесту та тесту на основі STACK за методологією CTT і IRT.	27.10.2025 – 02.11.2025	Виконано
6	Порівняльний аналіз стандартних покрокових тестів та тестів на основі STACK за методологією CTT та IRT.	03.11.2025 – 23.11.2025	Виконано
7	Аналіз проробленої роботи, підбиття підсумків та написання висновків.	24.11.2025 – 07.12.2025	Виконано
8	Оформлення магістерської дисертації.	08.12.2025 – 12.12.2025	Виконано

Студент Вікторія МИКИТЕНКО

Науковий керівник Ігор ОРЛОВСЬКИЙ

РЕФЕРАТ

Магістерська дисертація містить 80 сторінок, 27 рисунків, 10 таблиць, 1 додаток, 36 джерел літератури.

Робота присвячена аналізу якості покрокових комп'ютерних тестових завдань з математичних дисциплін, що використовуються в умовах дистанційного та змішаного навчання. Актуальність дослідження зумовлена необхідністю об'єктивного оцінювання не лише кінцевого результату, а й логіки розв'язування задач студентами, а також недостатньою кількістю психометричних досліджень покрокових тестів.

Метою роботи є аналіз структури покрокових тестових завдань з математичних дисциплін та оцінювання їхньої якості із застосуванням методів класичної теорії тестів (CTT) і моделей сучасної теорії тестування (IRT). Для досягнення поставленої мети здійснено аналіз наукових джерел з теорії тестування, досліджено структуру покрокових тестів, виконано ручний збір результатів тестування студентів і підготовку емпіричних даних до статистичного аналізу, а також проведено оцінювання психометричних характеристик тестових завдань методами CTT та IRT.

Об'єктом дослідження є покрокові комп'ютерні тести з математичних дисциплін університетського рівня, а предметом – психометричні характеристики якості покрокових тестових завдань.

У роботі використано методи аналізу та узагальнення наукової літератури, методи збору й попередньої обробки емпіричних даних, методи описової статистики, методи CTT та моделі IRT. Отримані результати дозволили оцінити надійність, дискримінативність і відповідність тестових завдань емпіричним даним, а також здійснити порівняльний аналіз стандартних покрокових тестів і тестів, реалізованих із використанням модуля STACK.

Практичне значення роботи полягає в можливості використання отриманих результатів для підвищення якості покрокових тестових завдань та удосконалення процесу оцінювання навчальних досягнень студентів з математичних дисциплін.

Ключові слова: *покрокові тести, комп'ютерне тестування, класична теорія тестів, сучасна теорія тестування, Moodle, STACK, graded response model, надійність і валідність тестів.*

Abstract

The master's thesis contains 80 pages, 27 figures, 10 tables, 1 appendix, 36 primary references.

The thesis focuses on the analysis of the quality of step-by-step computer-based tests in mathematical disciplines used in distance and blended learning environments. The relevance of the study is due to the need for objective assessment not only of results but also of students' problem-solving processes, as well as by the lack of psychometric studies of step-by-step tests.

The thesis aims to analyze the structure of step-by-step tests in mathematical disciplines and to evaluate their quality using methods of Classical Test Theory (CTT) and models of Item Response Theory (IRT). To achieve this aim, the study includes an analysis of scientific literature on test theory, an examination of the structure of step-by-step tests, manual collection of students' test results, and preparation of empirical data for statistical analysis, as well as an evaluation of the psychometric characteristics of test tasks using CTT methods and IRT models.

The object of the research is step-by-step computer-based tests in university-level mathematical disciplines, while the subject of the research is the psychometric characteristics of the quality of step-by-step test tasks.

The study employs methods of analysis and synthesis of scientific literature, methods of collecting and preprocessing empirical data, descriptive statistical methods, CTT methods, and IRT models. The results obtained made it possible to assess the reliability, item discrimination, and model fit of the test items to empirical data, as well as to conduct a comparative analysis of standard step-by-step tests and tests implemented using the STACK module.

The practical significance of the thesis lies in the possibility of using the obtained results to improve the quality of step-by-step tests and to enhance the assessment of students' learning outcomes in mathematical disciplines.

Keywords: *step-by-step tests, computer-based testing, Classical Test Theory, Item Response Theory, Moodle, STACK, graded response model, test reliability and validity.*

Зміст

Вступ

Стрімкий розвиток цифрових технологій суттєво змінив підходи до організації освітнього процесу в закладах вищої освіти. І в той же час дистанційне навчання перестало бути тимчасовою реакцією на надзвичайні обставини та перетворилося на невід'ємну складову сучасної освіти. Його ефективність залежить не лише від наявності технологічних платформ, але й від методично обґрунтованих підходів до викладання та оцінювання. Особливо гостро ці питання постають у математичній освіті, де важливо оцінювати не тільки кінцевий результат, а й логіку, послідовність та коректність міркувань студента.

Одним з найважливіших інструментів сучасного дистанційного навчання є комп'ютерне тестування. Тестові методи оцінювання дозволяють вимірювати латентні характеристики студентів: рівень підготовки, успішність та сформованість математичних компетентностей за об'єктивними та стандартизованими критеріями. Використання електронних систем оцінювання забезпечує неупередженість, однакові умови контролю, можливість швидко перевіряти великі групи студентів і значно зменшує навантаження на викладача. З огляду на це, питання створення якісних тестових інструментів постійно перебуває в центрі уваги педагогів і дослідників, серед яких варто згадати Г. Раша, М. Вільсона, В.С. Аванесова, Р.К. Хемблтона, В.Я. Ван дер Ліндена, а також українських учених Л.О. Кухар, Т.В. Лісову та Ю.О. Ковальчука.

З-поміж численних платформ особливе місце займає система Moodle, яку використовують понад 150 000 навчальних закладів у світі. Moodle надає широкий спектр інструментів для самостійної роботи студентів: структуровані курси, інтерактивні завдання, вбудовані механізми тестування. Проте ефективність платформ залежить не лише від доступу до них, але й від грамотної методики побудови завдань, які здатні перевіряти повний хід розв'язання математичних задач.

Традиційні тести (множинний вибір, коротка відповідь) часто не дозволяють коректно оцінити глибину розуміння матеріалу. Тому дедалі більше уваги приділяється розробці більш складних тестових завдань. Одним із таких рішень є покрокові тестові завдання, які дають змогу фіксувати проміжні кроки міркування, перевіряти алгоритм розв'язання, а не лише остаточний результат. Перевагою покрокових тестів є їхня максимальна наближеність до письмових контрольних робіт за умови повної автоматизації оцінювання.

Ще одним перспективним інструментом є плагін STACK, який дозволяє студентам демонструвати міркування, вводити математичні вирази у символьній формі та отримувати детальний зворотний зв'язок. STACK підвищує об'єктивність перевірки та надає цінну діагностичну інформацію щодо помилок і прогалин у знаннях.

У сучасних умовах важливим є не лише розробка подібних завдань, а й забезпечення їхньої психометричної якості. Методологічні підходи класичної теорії тестування та сучасної теорії тестування дозволяють аналізувати надійність, валідність, складність і дискримінативність тестів, визначати ефективність окремих завдань та порівнювати різні їх формати.

З огляду на зазначене, актуальним є комплексне дослідження структури та якості покрокових тестів і тестів на основі STACK у математичній освіті університетського рівня.

Метою даної магістерської роботи є дослідження специфіки побудови покрокових тестових завдань та аналіз їхньої якості з використанням методів класичної та сучасної теорії тестування, а також порівняння ефективності тестів, реалізованих за допомогою стандартних можливостей Moodle, з тестами, побудованими за допомогою плагіна STACK.

Розділ 1. Класична теорія тестів та сучасна теорія тестів

1.1 Вступ

Оцінювання результатів навчання студентів є дуже важливим в освіті. У сучасній педагогічній практиці оцінювання когнітивних здібностей, академічних навичок та інтелектуального розвитку здобувачів освіти передбачає використання певних методик, що спрямовані на вимірювання результатів навчальної діяльності студентів як за всім курсом загалом, так і за певною темою зокрема. Найпоширенішою такою методикою є тестування, яке дає змогу не лише простежити рівень знань студентів, а і виявити закономірності у їхніх відповідях та відповідно вдосконалити навчальний процес.

Таким чином, створення якісних тестів є дуже важливим для коректної оцінки успішності учнів, саме тому важливо перевіряти його психометричні характеристики і враховуючи отримані результати покращувати розроблені завдання. Серед найвідоміших підходів для оцінювання якості тестів – класична теорія тестів (англ. Classical Test Theory, далі використовуватимемо скорочення CTT) та сучасна теорія тестів (англ. Item Response Theory, далі використовуватимемо скорочення IRT. Дослівний переклад українською мовою як «теорія відповідей на питання» звучить примітивно і не відображає суті даної теорії, тому найчастіше в україномовній літературі цю теорію називають сучасна теорія тестів на відміну від класичної теорії).

Класична теорія тестування є одним з найважливіших та найпростіших розділів психометрії, що використовується для побудови та оцінки психологічних і педагогічних тестів. Основною його метою є прогнозування загальних результатів тестів або відповідей на окремі тестові завдання на основі вже завершених тестів.

Цей підхід почав своє існування у першій половині XX-го століття як теоретична основа, спрямована на пояснення результатів тестування та оцінювання у простій манері, спираючись всього лише на два основні компоненти: справжня оцінка та спостережувана оцінка. Відповідно дана

теорія дозволяє пояснити індивідуальну успішність студента на основі різниці між цими двома балами. Вона припускає, що справжній бал індивіда відображає його фактичний рівень здібностей або знань, тоді як спостережуваний бал може містити частину похибок, що виникають внаслідок випадкових факторів, які можуть вплинути на вимірювання, таких як умови навколишнього середовища індивіда під час проходження тесту або труднощі з розумінням інструкцій.

Головною метою CTT є підвищення надійності тестів, зазвичай двома способами: перший – використання надійності повторного тестування (test-retest reliability) – тестове завдання вважається надійним, якщо тестувальники повторюють свою відповідь протягом кількох різних спроб; другий – альтернативна надійність тестування (alternate test reliability) – тестове завдання вважається надійним, якщо тестувальники повторюють свою відповідь в альтернативній версії того самого тесту. Таким чином, аналіз зібраних даних у межах цієї теорії дає змогу оцінити, наскільки тест є послідовним і передбачуваним у різних вибірках.

Хоча CTT була значним проривом у галузі психологічних вимірювань, з швидким розвитком науки й техніки виявилися її певні обмеження. Серед них можна виокремити одне найважливіше: значна залежність результатів аналізу якості тесту від конкретної вибірки респондентів. Крім того, класичній теорії тестування бракує гнучкості у роботі з завданнями різної складності або дискримінативності, що спонукало дослідників шукати більш просунуті моделі, такі як, наприклад, сучасна теорія тестування, що замінила CTT у багатьох сучасних психологічних та освітніх контекстах. На противагу своїй попередниці, ця теорія аналізує кожне завдання незалежно одне від одного, а також спирається на складніші математичні моделі для кращої інтерпретації відповідей студентів на основі їхніх фактичних здібностей та складності завдань.

1.2 Основи класичної теорії тестування

Як вже було зазначено раніше, класична теорія тестування пропонує простий та чіткий спосіб розуміння результату навчальної діяльності студента, адже згідно з цією теорією, спостережуваний бал індивіда складається із суми істинного балу та випадкової похибки [1]. Цю концепцію можна представити наступним рівнянням

$$X = T + E,$$

де X – спостережуваний бал, T – істинний бал, а E – випадкова похибка.

При цьому передбачається, що похибка оцінки з тестування має однаковий розподіл у межах усієї групи учасників тестування [11]. Інакше кажучи, незалежно від того, чи має студент високий, середній або низький результат, рівень стандартної похибки вимірювання його здібностей залишається сталим.

Це рівняння підкреслює простоту моделі, але водночас і висвітлює обмеження теорії у врахуванні зовнішніх чинників. Наприклад, екологічні або психологічні фактори можуть збільшити частку випадкової похибки, тим самим знижуючи точність оцінок [2].

Основними показниками класичної теорії тестування є надійність та валідність.

Принцип надійності базується на здатності тесту давати стабільні результати при його повторному застосуванні за тих самих умов. Вона розраховується за допомогою різних коефіцієнтів, проте одним з найрозповсюдженіших є альфа Кронбаха, який використовується для визначення внутрішньої узгодженості завдань в тесті [2].

Отже, надійність можна визначити як ступінь, до якого результат учасника тесту не залежить від: дня та часу проведення тестування або місця перебування тестувальника; конкретних запитань або задач, які були в одній з варіацій тесту, що складав учасник, але яких немає в поточному тесті;

конкретних оцінювачів, що виставляли рейтинг учасника тесту (якщо процес оцінювання включав будь-яке людське судження).

Надійність є важливим елементом будь-якого тесту та вважається ключовим показником його якості. Однак дослідники можуть постати перед труднощами у підтримці високого рівня надійності під час застосування тестів у різних середовищах або на різноманітних вибірках. Наприклад, рівень надійності може змінюватися, коли тест проводиться для груп населення, які відрізняються культурою або віком, що може вимагати повного перероблення тесту або модифікації деяких завдань [4]. Отже, надійність у CTT не завжди зберігається за зміни умов, що є одним із її практичних обмежень.

Валідність – це ще один ключовий аспект якості тесту. Без валідності тест втрачає свою цінність як інструмент вимірювання, адже вона показує ступінь, до якого тест вимірює те, що він повинен вимірювати. У CTT валідність оцінюється за допомогою різних методів, таких як загальна валідність тесту та валідність окремого завдання. Однак, не зважаючи на те, що класична теорія тестування надає потужні інструменти для вимірювання валідності, існують певні обмеження в їх застосуванні, особливо в тестах, які повинні передбачати індивідуальну ефективність студентів у різних контекстах [2]. Крім того, іноді може бути складно точно виміряти валідність у тестах, що стосуються абстрактних або багатовимірних понять.

Отже, поняття надійності та валідності дійсно схожі в деяких важливих аспектах, проте є суттєво різними в інших. Надійність спирається на пошук конкретного джерела невідповідності в результатах тесту, а валідність перевіряє ціль конкретного використання тесту: вона залежить від того, як використовуються результати тесту, в той час як надійність – ні.

Однак, не зважаючи на відмінності, обидва ці показники є ключовими для оцінки якості тестів: адже, якщо ви вимірюєте неправильну річ, неважливо, наскільки добре ви її вимірюєте, але водночас якщо результати вимірювань сильно залежать від випадковості, ви насправді нічого не вимірюєте.

1.3 Основні недоліки класичної теорії тестування

Не дивлячись на широке використання класичної теорії тестів у різних психологічних та освітніх галузях, вона не обходиться без критики. Існує кілька основних недоліків та проблем, пов'язаних із застосуванням цієї теорії для вимірювання рівня знань студентів.

Найважливішим недоліком є значна залежність цієї теорії від характеристик вибірки, що використовується при побудові та оцінці якості тесту. Це означає, що результати самого тесту та його основні показники (такі як надійність та валідність) можуть бути обмежені групою, що проходила тест, оскільки СТТ не враховує належним чином індивідуальні відмінності студентів [7]. Тобто якщо тест створено для вузької вибірки (наприклад, групи сильних студентів), його результати можуть не відтворюватися в інших групах. Це в свою чергу означає, що розробка тесту з використанням СТТ вимагає ретельного відбору та різноманітності вибірок для забезпечення узагальнюваності результатів.

Також згідно класичної теорії тестування, всі елементи тесту мають однакову вагу, а також припускається, що кожне завдання робить однаковий внесок у загальний бал тесту. Однак ці припущення можуть не справджуватися, особливо коли йдеться про багатовимірні тести або завдання, які суттєво відрізняються за складністю або здатністю розрізняти здібності окремих осіб [8]. Це і є найбільшим недоліком класичної теорії тестування, де всі елементи розглядаються однаково. У складних тестах це може призвести до неточних результатів, оскільки деякі завдання можуть мати більший вплив на загальний рівень знань окремих осіб. Наприклад, одне дуже складне завдання може непропорційно визначати загальний бал людини, що робить оцінку її здібностей неточною. Ця критика підкреслює необхідність більш просунутих моделей, таких як сучасна теорія тестування, яка вже враховує різницю у складності завдань.

Як зазначалося раніше, СТТ спирається на припущення, що тест проводиться в ідеальних та стабільних умовах. Однак насправді зовнішні фактори, такі як загальний стан довкілля, час та місце проведення можуть суттєво впливати на результати тестування студентів (особливо якщо ми говоримо про сьогоднішні реалії українців). Наприклад, невідповідні навколишні умови, такі як зайвий шум або стрес, можуть вплинути на результати тестування студента, що призведе до неточних результатів [2]. Зовнішні умови є фактором, який важко контролювати в психологічних вимірюваннях. СТТ значною мірою ігнорує ці впливи, а це означає, що будь-який вплив довкілля чи психології може призвести до неточних або оманливих результатів.

Ну і наостанок варто зазначити, що однією з основних проблем СТТ є її обмеження в оцінці справжніх здібностей людини. Адже оскільки вона спирається на порівняння спостережуваного балу зі справжнім, будь-яка похибка вимірювання безпосередньо впливає на точність цих оцінок [2].

Очевидно, що надання точної оцінки справжніх здібностей студента є фундаментальною метою будь-якого психологічного вимірювання. Однак, СТТ значною мірою спирається на спрощені припущення, які не розрізняють індивідуальні відмінності респондентів. Як наслідок, повна залежність від СТТ може призвести до неточних або неефективних тестів для оцінки справжніх здібностей, що зменшує саму цінність результатів тестів. Отже, розробникам необхідно знати про ці обмеження при використанні СТТ для розробки інструментів вимірювання.

1.4 Основи сучасної теорії тестів

Сучасна теорія тестування передбачає використання математичних функцій для прогнозування успішності студента в тесті за допомогою набору факторів, які називаються прихованими здібностями (невидимі риси, які тест має оцінити через відповіді на конкретні завдання). Це теорія тестування, яка ґрунтується на взаємозв'язку між успішністю осіб на окремому тестовому

завданні та рівнем успішності студентів за загальним показником здібності, для оцінювання якого це завдання було розроблено [2].

IRT базується на ідеї, що ймовірність правильної відповіді на завдання є математичною функцією параметрів респондента (рівень його здібностей) і завдання (складність, дискримінативність та схильність до вгадування). Цей зв'язок між успішністю учня в тестових завданнях та показником його здібностей можна описати за допомогою так званої функції відгуку завдання. Для простих завдань, таких як правильні чи неправильні відповіді, сучасна теорія тестування вказує, що ймовірність правильної відповіді на завдання повинна зростати передбачуваним чином зі зростанням рівня здібностей учня. Для завдань з множинним вибором необхідно використовувати складніші моделі теорії відповідей на завдання.

Варто також зазначити, що при використанні сучасної теорії тестування для створення тестових завдань робиться припущення, що існує один домінантний фактор або здатність, яка може пояснити успішність у тесті. Здібність або риса, що вимірюється тестовим завданням, визначається широко або вузько з точки зору здібностей, досягнень або особистості.

Інструменти тестування сучасної теорії надають як інваріантну статистику завдань, так і оцінки здібностей. За допомогою процесу оцінки параметрів тестові завдання та учні розміщуються на шкалі здібностей таким чином, щоб існував якомога тісний зв'язок між очікуваними ймовірностями успіху учнів у тестових завданнях, отриманими на основі розрахованих моделлю параметрів завдання та здібностей, та фактичними результатами учнів на кожному рівні здібностей. Ці оцінки параметрів завдань та оцінки здібностей учнів постійно переглядаються, доки не буде досягнуто найкращої можливої узгодженості між прогнозами, заснованими на оцінках здібностей та параметрів завдання і фактичними даними тесту.

Отже, сучасна теорія тестування пропонує вдосконалену альтернативу класичній теорії тестування, яка характеризується здатністю забезпечувати точніший аналіз та враховувати індивідуальні здібності респондентів. У той

час як CTT зосереджується на аналізі загальної результативності тесту на основі спостережуваного балу, IRT наголошує на аналізі кожного завдання окремо з точки зору складності та дискримінативності, пропонуючи точні оцінки здібностей індивідів на основі їхніх відповідей на різні завдання. IRT використовує математичні моделі для аналізу індивідуальних відповідей на кожне завдання незалежно, що допомагає краще оцінити справжні здібності індивідів порівняно з традиційними моделями.

1.5 Переваги та недоліки сучасної теорії тестів

Сучасна теорія тестів має низку переваг у порівнянні з класичною теорією тестування. Насамперед, вона відрізняється здатністю забезпечувати точніші оцінки здібностей індивідів порівняно з CTT. Зосереджуючись на аналізі кожного завдання окремо, ця теорія здатна запропонувати чіткіше уявлення про результативність індивіда та його справжні здібності. Одна з ключових переваг сучасної теорії тестів полягає в її здатності аналізувати рівень складності кожного завдання, і таким чином оцінювати фактичну продуктивність тесту на основі відповідей людей на завдання різної складності. Також IRT забезпечує більшу гнучкість порівняно з CTT, оскільки може обробляти багатовимірні тести або тести, що містять значні відмінності у складності завдань [4]. Ця функція дозволяє викладачам отримувати більш точні оцінки здібностей студентів, які є значно менш залежними від статистичного розподілу вибірки. Наприклад, у тесті, який містить завдання від легких до складних, IRT може точніше визначити індивідуальні здібності на основі того, як людина реагує на більш складні завдання порівняно з простішими.

Однією з найважливіших переваг IRT є те, що її моделі не сильно залежать від характеристик вибірки, на відміну від CTT. В даній теорії питання аналізуються на основі відповідей окремих осіб незалежно від використаної вибірки, що робить результати більш узагальнюваними для інших популяцій [3]. Ця особливість робить IRT надійною моделлю, особливо при застосуванні

тестів у різноманітних середовищах. А отже, на відміну від класичної теорії тестування, яка вимагає великої репрезентативної вибірки для узагальнення результатів, IRT дозволяє використовувати тести на різних вибірках без суттєвого впливу на результати. Це робить сучасну теорію тестів більш точним інструментом для вимірювання при роботі з різними групами, тим самим сприяючи підвищенню точності психологічних та освітніх оцінок.

Проте, попри численні переваги, сучасна теорія тестування має і ряд недоліків. Очевидно, через свої припущення IRT є складнішою в аналізі та впровадженні порівняно з CTT, адже вимагає використання передових математичних моделей для аналізу даних, а це означає, що дослідникам потрібне спеціалізоване програмне забезпечення та глибоке розуміння статистичної математики для проведення аналізу [7]. Отже, хоча IRT і забезпечує більшу точність, але пов'язана з нею складність створює перешкоду для багатьох дослідників та установ, які можуть не мати необхідних ресурсів або досвіду для використання цієї моделі. У багатьох випадках вартість придбання спеціалізованого програмного забезпечення може бути високою, що робить CTT кращим варіантом у середовищах, які вимагають простого та швидкого аналізу. Як наслідок, незважаючи на теоретичну перевагу даної теорії, ця складність обмежує широке впровадження IRT.

Але разом з тим, одним з найголовніших обмежень IRT є підвищені вимоги до обсягу емпіричних даних для забезпечення точності оцінок. Аналіз цієї теорії спирається на надання точної інформації про складність та дискримінативність кожного завдання тесту, що вимагає достатнього розміру вибірки для отримання надійних оцінок. Це створює труднощі, особливо в дослідженнях з обмеженими вибірками, або в практичних тестових контекстах, таких як індивідуалізовані освітні тести [5]. У таких випадках використання методології CTT може бути більш доцільним для первинного аналізу, оскільки вона не висуває настільки суворих вимог до розміру вибірки.

Отже, незважаючи на перевагу IRT у наданні точніших оцінок здібностей людей та подолання деяких обмежень, присутніх у CTT,

складності, пов'язані з його впровадженням та необхідністю великих вибірок, роблять його інструментом, який важко використовувати в широких масштабах.

1.6 Порівняння CTT та IRT

Класична та сучасна теорія тестування представляють дві різні моделі для аналізу та оцінки результатів у психологічних та освітніх тестах. У той час як CTT зосереджується на середніх значеннях, надійності та валідності та спирається на аналіз загальних спостережуваних балів, IRT зосереджується на детальному аналізі кожного завдання окремо, враховуючи його складність та дискримінативність, і на індивідуальних здібностях тестувальників.

Хоча CTT може бути більш поширеною та легшою у застосуванні, IRT забезпечує набагато більшу точність в оцінці індивідуальних здібностей та параметрів завдання. З точки зору точності, IRT дозволяє детальніше оцінювати кожен пункт окремо, що дає змогу дослідникам виявити дрібні індивідуальні відмінності між респондентами. Однак, з точки зору необхідних ресурсів, IRT вимагає додаткових витрат та складнішого аналізу порівняно з CTT.

Отже, класична теорія тестів залишається придатною у ситуаціях, які не потребують глибокого аналізу, або під час роботи в умовах обмежених ресурсів, тоді як IRT перевершує її в академічних або дослідницьких умовах, які вимагають високої точності оцінювання.

З огляду на зазначене, у даній роботі використовується поєднання методів CTT та сучасних моделей IRT, що дозволяє здійснити комплексний аналіз якості покрокових тестових завдань та отримати більш ґрунтовні результати.

Розділ 2. Методологія оцінювання якості тестів за CTT

Даний розділ присвячений оцінюванню якості тестових завдань. Підходи, представлені тут, були розроблені в рамках теоретичної бази CTT. Отже, для початку будемо вважати, що тест складається з низки завдань і вже був проведений серед певної вибірки студентів. Після того як студенти завершили тест, можна розпочинати аналіз його якості. Під час дослідження якості математичних тестів зазвичай використовується комбінація психометричних показників класичної теорії тестування та звичайних вибіркових характеристик математичної статистики.

Далі пропоную детальніше розглянути кожен з цих показників, зокрема їхні означення, математичні формули, інтерпретацію отриманих результатів та обґрунтування включення до аналізу якості тесту. Більш детальну інформацію про кожну з описаних нижче характеристик тесту можна знайти у джерелах [14;30-35].

Середній бал відображає середню успішність учнів з певного завдання або з тесту в цілому та визначається наступною формулою

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

де x_i позначає бал i-го учня, а n – загальну кількість респондентів.

Середнє значення близьке до 1 вказує на те, що завдання є дуже легким, тоді як середнє значення близьке до 0 означає, що завдання є дуже складним. В якісному тесті середнє значення має бути в межах від 0.4 до 0.8, що забезпечує достатню складність та потенціал для розпізнавання.

Медіана – це середній елемент для набору оцінок за тест, упорядкованих за зростанням або спаданням. На медіану менше впливають викиди та перекошені дані. Щоб її обчислити, спочатку потрібно побудувати варіаційний ряд вибірки загальних оцінок за тест або за окреме завдання $x_{(1)} < x_{(1)} < \cdots < x_{(n)}$, далі

$$M_e = \begin{cases} x_{(k+1)}, & \text{якщо } n = 2k+1; \\ \dfrac{x_{(k)} + x_{(k+1)}}{2}, & \text{якщо } n = 2k. \end{cases}$$

Якщо медіана сильно відрізняється від середнього значення, це вказує на асиметрію в розподілі балів.

Мода – це значення елемента вибірки, яке найчастіше зустрічається у ній. В нашому випадку будемо використовувати формулу на основі інтервального розподілу

$$M_0 = x_0 + h \cdot \frac{f_{m_0} - f_{m_{0-1}}}{\left(f_{m_0} - f_{m_{0-1}}\right) + \left(f_{m_0} - f_{m_{0+1}}\right)},$$

де x_0 – нижня межа модального інтервалу, h – ширина модального інтервалу, $f_{m_0}, f_{m_{0-1}}, f_{m_{0+1}}$ – частоти відповідно модального, передмодального та післямодального інтервалів.

На стовпчастій гістограмі мода представляє найвищий стовпчик. Зазвичай мода використовується лише для категоріальних даних, де ми хочемо знати, яка категорія найчастіше зустрічається у вибірці. Якщо мода сильно відрізняється від середнього значення, це може свідчити про нерівномірний розподіл.

Стандартне відхилення вказує на ступінь мінливості відповідей учнів, обчислити його можна за формулою

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

де x_i позначає бал i-го учня, $\bar{x}$ – середнє значення балу у вибірці, а n – загальну кількість респондентів.

Велике стандартне відхилення свідчить про різноманітну успішність студентів, що бажано для розрізнення рівнів здібностей учнів. Низьке SD, навпаки, сигналізує про однорідність відповідей та занижену діагностичну здатність тесту або завдання.

Коефіцієнт асиметрії – числова характеристика розподілу ймовірностей дійсної випадкової величини. Знаходимо його за формулою

$$sk = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sqrt[3]{SD}}.$$

Якщо коефіцієнт асиметрії близький до нуля – це вказує на симетричність розподілу оцінок за тест і відповідно його низьку діагностичну здатність. Від'ємна асиметрія вказує на відносно легке завдання або тест («хвіст» розподілу оцінок за тест довший ліворуч, тобто ми маємо багато високих балів в тесті), тоді як додатна асиметрія відповідає складнішому завданню або тесту («хвіст» розподілу довший праворуч, багато низьких балів).

Коефіцієнт ексцесу оцінює концентрацію балів навколо середнього значення, його зазвичай розраховують за наступною формулою

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{SD^4} - 3.$$

Додатне значення ексцесу вказує на піковий розподіл з більшістю балів поблизу середнього значення (лептокуртичний), тоді як від'ємне значення вказує на більш пологий, розсіяний розподіл (платикуртичний). Цей параметр особливо корисний для виявлення аномальних закономірностей у виконанні завдань.

Кореляція завдання-тест (Item-Total Correlation, ITC) представляє кореляцію Пірсона між балами за певне завдання та загальним балом за всі тестові завдання

$$ITC_j = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (x_{ij} - \bar{x_j})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(x_{ij} - \bar{x_j})^2}},$$

де x_i – бал i-го учня за весь тест, x_{ij} – бал, який отримав і-ий студент за j-те завдання, $\bar{x}$ – середній бал за весь тест, $\bar{x_j}$ – середній бал за j-те завдання, n – загальна кількість респондентів.

При значенні $ITC_j \geq 0.7$ завдання тесту має високу роздільну здатність. Якщо значення $ITC_j < 0.7$ це означає, що завдання потребує доопрацювання. Від'ємні значення ITC_j свідчать про слабку відповідність із загальною конструкцією тесту і вказують на неправильність складання вказаного завдання. Такі завдання слід виключати із тесту.

Кореляція завдання - решта тесту (Item-Rest Correlation, IRC) – це кореляція між балом за окреме завдання та сумарним балом за тест без урахування оцінки за це завдання (тобто решти тесту). Цей показник використовується для оцінки розділювальної здатності завдання: наскільки добре воно узгоджується з рештою тесту. IRC зазвичай використовується як більш надійний індикатор при розробці тестів, адже ITC може бути трохи завищеним, оскільки саме досліджуване завдання включено в загальний бал, а от IRC є вже більш «чистим» індикатором. Якщо обидва коефіцієнти низькі – це чіткий сигнал для перегляду завдання. Всі розрахунки та інтерпретація цього показника аналогічні до кореляції завдання-тест.

Також для оцінки якості тесту можна побудувати ***кореляційну матрицю завдань*** та проаналізувати її. Для того щоб тестові завдання не повторювали одне одного, значення кореляції між ними має бути не високим $r_{ij} \leq 0.3$. Якщо ж дане завдання має від'ємну кореляцію з більшістю інших завдань, то необхідно зробити заміну цього завдання.

У процесі розробки та аналізу тестів існує так званий парадокс кореляцій завдання-завдання та завдання-тест. Його суть полягає в тому, що завдання тесту, які забезпечують високу внутрішню узгодженість, не завжди сприяють високій валідності щодо зовнішніх критеріїв. Якщо тест створений, щоб вимірювати однорідний конструкт, його завдання будуть тісно пов'язані між собою – матимуть високі кореляції завдання-завдання. Однак у цьому разі кожне завдання може не мати сильної кореляції із зовнішнім критерієм, якщо критерій охоплює кілька різних аспектів здібностей студента. І навпаки, якщо тест спеціально розроблений для оцінювання різних аспектів складного,

багатовимірного критерію, його пункти будуть менш взаємопов’язаними між собою. Проте в такому випадку всі тестові завдання можуть мати високу кореляцію із зовнішнім критерієм, оскільки всі вони охоплюють окремі його компоненти. Отже, важливо завжди чітко розуміти мету використання тесту. Якщо застосовувати тестові бали не відповідно до логіки, закладеної у його розроблення, це може призвести до парадоксальних або хибних результатів оцінювання.

Отже, повертаючись до кореляції завдання-завдання, рекомендації на основі отриманих коефіцієнтів наступні:

1. Низька кореляція між завданнями (< 0.3): завдання погано «вписується» в тест, воно може вимірювати іншу здібність або мати проблеми з формулюванням. Потрібно перевірити зміст, інструкцію або статистичну складність завдання.
2. Помірна кореляція між завданнями ($0.3 - 0.7$): ідеальний рівень – всі завдання різні, але вимірюють той самий психологічний конструкт. Такі пари завдань підвищують внутрішню узгодженість без надлишкової кількості завдань.
3. Дуже висока кореляція між завданнями (> 0.7): завдання можуть бути надмірно схожими або навіть дублюючими. Це знижує ефективну довжину тесту: тест стає довшим без приросту нової інформації. Потрібно видалити одне із завдань або переробити його.
4. Від’ємна кореляція: можлива помилка в ключі, змістова відмінність або інверсія формулювання. Такі завдання майже завжди потребують перегляду.

Коефіцієнт середньої кореляції між завданнями (Average Item-Item Correlation, AIC) використовується для оцінки внутрішньої узгодженості тесту, показуючи, наскільки окремі питання тесту є взаємопов’язаними між собою. Формула середньої кореляції між завданнями

$$AIC = \frac{2}{n(n-1)} \sum_{i \neq j}^{n} r_{ij} ,$$

де r_{ij} – парні кореляції між усіма питаннями тесту, n – загальна кількість питань у тесті.

Інтерпретація коефіцієнта середньої кореляції зазвичай наступна: $AIC \in [0.15\,;0.50]$ – оптимальний рівень кореляції (тест добре збалансований); $AIC < 0.15$ – низька кореляція між питаннями (може вказувати на неоднорідний тест, тобто завдання в тесті вимірюють різні конструкції – бажано провести факторний аналіз); $AIC > 0.50$ – надмірна кореляція (можливе дублювання питань – надто схожі питання варто або прибрати з тесту, або переформулювати).

Індекс легкості завдання в класичній теорії тестування є першою характеристикою завдання, яку потрібно визначити. Все дуже просто: якщо оцінка за завдання є неперервною в певному інтервалі, то легкість завдання – це середня оцінка за дане завдання. Чим вищий середній бал за дане завдання, тим легшим воно є. Формула для індексу легкості завдання наступна

$$p = \frac{R}{M} ,$$

де R – середня оцінка за дане завдання тесту, а M – максимальна можлива кількість балів за це завдання.

Отже, який оптимальний рівень p для завдання тесту? Завдання з рівнями легкості 1 або 0 є марними, оскільки вони не розрізняють окремих студентів. Тобто, якщо всі здають завдання, це діє так само, як додавання 1 до загального балу кожного. Якщо всі не здають завдання, то до балу кожного додається 0. Час, витрачений на написання завдання, його створення, відповідь на нього та оцінювання, витрачається даремно.

Завдання зі значеннями індексу 0.50 забезпечують найвищі рівні диференціації між особами в групі. Єдине застереження щодо значення індексу 0.50 як найкращого виникає, коли завдання мають високу

взаємокореляцію. Якщо це так, то ті самі 50% респондентів складуть всі завдання, і одного завдання, а не всього тесту, було б достатньо, щоб розмежувати учасників тестування на дві групи.

Враховуючи все це, загальні рекомендації щодо інтерпретації індексу легкості завдання наступні: якщо індекс легкості менше за 0.3, то завдання вважається дуже складним, значення індексу в межах від 0.3 до 0.7 вважаються оптимальними, якщо ж легкість завдання більша за 0.7 – завдання дуже легке [12].

Індекс дискримінативності завдань можна застосувати для обчислення простої міри розрізнювальної здатності завдання використовуючи метод крайніх груп. При розрахунку індексу D вибірку з результатами учнів спочатку впорядковують в порядку зростання. Далі для аналізу відокремлюють 27% учнів з найвищими результатами та 27% з найнижчими. Використовується саме 27% квантиль, оскільки показано, що саме це значення максимізує відмінності в нормальних розподілах, забезпечуючи при цьому достатньо випадків для аналізу [13]. Якщо оцінка за завдання це неперервне значення в певному діапазоні, то індекс дискримінативності завдання обчислюється наступним чином

$$D = P_u - P_l \,,$$

де P_u – середня оцінка за дане завдання для верхньої групи, а P_l – середня оцінка за дане завдання для нижньої групи.

Очевидно, що значення цього індексу коливаються від -1 до +1, при чому від'ємне значення означає, що більша частина нижньої групи відповідала на завдання правильно, тоді як додатний індекс означає, що більша частка верхньої групи відповідала на завдання правильно.

Значення індексу вище за 0.4 зазвичай свідчить про хорошу дискримінативну здатність завдання, якщо ж індекс знаходиться в межах від 0.3 до 0.4, то це означає, що було б непогано дане завдання трошки удосконалити, якщо індекс потрапляє в межі від 0.2 до 0.3, то це вказує на

необхідність перегляду даного завдання, при індексу менше 0.2, завдання потрібно або виключити, або повністю переробити.

Коефіцієнт надійності тесту альфа Кронбаха вперше була описана у статті 1951 року і з того часу є показником внутрішньої узгодженості тестів, що найчастіше використовується. Цей показник оцінює ступінь, до якого елементи тесту послідовно вимірюють одну й ту саму базову конструкцію.

Цей коефіцієнт розраховується як

$$\alpha = \frac{N}{N-1}\left(1 - \frac{\sum_{i=1}^{N} \delta_i^2}{\delta_T^2}\right),$$

де N – кількість тестових завдань, δ_i^2 – дисперсія окремих завдань, а δ_T^2 – загальна дисперсія балів.

Інтерпретація коефіцієнта α Кронбаха наступна: $\alpha \geq 0.90$ – відмінна надійність; $0.80 \leq \alpha < 0.90$ – добра надійність; $0.70 \leq \alpha < 0.80$ – прийнятна надійність; $0.60 \leq \alpha < 0.70$ – сумнівна надійність; $\alpha < 0.60$ – погана надійність.

Проте варто зазначити, що альфа Кронбаха базується на припущенні, що відповіді на окремі запитання розподілені нормально, мають однакову дисперсію та однаково пояснюють загальний фактор. Ці припущення рідко повністю виконуються в реальних умовах, тому альфа буде непридатною та надаватиме неточні або упереджені оцінки внутрішньої узгодженості. Як наслідок, тести, які насправді мають високу внутрішню узгодженість, можуть бути відхилені або модифіковані, або тести, які насправді мають низьку внутрішню узгодженість, можуть бути збережені.

Омега Макдональда – це міра надійності для тестів та анкет, розроблена психологом Дональдом Р. Макдональдом. Омега Макдональда базується на факторно-аналітичному підході, на відміну від альфи, який в основному базується на кореляції між питаннями. Омега виявилася більш стійкою, ніж альфа, до відхилень від припущень, зазначених вище, і тому загалом буде більш підходящим показником внутрішньої узгодженості.

Вона розраховується як

$$\omega = \frac{(\sum_{i=1}^{N} \lambda_i)^2}{(\sum_{i=1}^{N} \lambda_i)^2 + \sum_{i=1}^{N} \theta_i}.$$

Формула базується на факторних навантаженнях, отриманих з моделі підтверджувального факторного аналізу (CFA), де λ_i представляє факторні навантаження тестових завдань, θ_i представляє дисперсії помилок для кожного завдання.

Також варто зауважити, що коефіцієнт надійності тесту омега Макдональда має дві основні варіації: ω_{total} (загальний) і $\omega_{\text{hierarchical}}$ (ієрархічний).

Загальний коефіцієнт омега показує, яка частка дисперсії загального тестового балу зумовлена усіма спільними факторами (і головним, і другорядними). Тобто якщо, наприклад, загальний коефіцієнт омега рівний 0.9, то це означає що 90% варіації загального балу за тест пояснюється спільними факторами, а 10% – шумом.

Ієрархічний коефіцієнт омега показує, яка частка дисперсії загального тестового балу зумовлена генеральним фактором (без будь-якого впливу другорядних факторів). Тобто якщо, наприклад, ієрархічний коефіцієнт омега рівний 0.9, то це означає що 90% варіації загального балу за тест пояснюється саме головним фактором, а 10% – другорядними факторами або шумом.

Якщо ми аналізуємо тест з математичних дисциплін, то факторами можна вважати тематичні або когнітивні компоненти знань студентів, що проявляються у зв'язках між завданнями (наприклад, якщо ми говоримо про комплексне тестування з математичного аналізу, то окремими підфакторами можуть бути «уміння диференціювати», «уміння інтегрувати», «уміння працювати з рядами», «знання функцій комплексної змінної», а от генеральним фактором можна назвати «загальне володіння математичним аналізом»).

Отже, загальний коефіцієнт омега показує, наскільки всі завдання тесту узгоджені в цілому, незалежно від того, скільки в ньому приховано підтем (факторів). Ієрархічний показує, наскільки ці завдання вимірюють один

спільний головний фактор (наприклад, «загальні знання з матаналізу»), а не лише окремі теми.

Інтерпретація ω Макдональда ідентична до інтерпретації альфа Кронбаха.

Оцінка нормальності розподілу балів тесту є важливим кроком в аналізі тестів, оскільки багато методів статистики (наприклад, t-тести, дисперсійний аналіз, регресійні моделі) спираються на припущення, що базові дані відповідають нормальному розподілу. Відхилення від нормальності можуть призвести до упереджених оцінок, неправильних висновків або зниження статистичної потужності. Для перевірки цього припущення можна застосовувати два додаткові статистичні тести: тест Шапіро-Вілка та тест Колмогорова-Смірнова.

Першим кроком для виконання тесту потрібно «нормувати» отримані індивідуальні бали студентів, тоді гіпотеза про нормальність даних вибірки прийме вигляд:

H_0: вибірка індивідуальних балів має стандартний гауссівський розподіл $N(0,1)$.

H_1: розподіл вибірки індивідуальних балів не є гауссівським.

Тест Колмогорова-Смірнова оцінює максимальну абсолютну різницю між емпіричною функцією розподілу вибірки та кумулятивною функцією розподілу заданого теоретичного розподілу – у нашому випадку нормального

$$D = \sup_x |F_n(x) - F(x)|,$$

де $F_n(x)$ – ЕФР вибірки, а $F(x)$ – КФР нормального розподілу.

Для проведення тесту спочатку потрібно обрати рівень значущості α. Далі необхідно на основі нормованих даних побудувати інтервальний статистичний розподіл, розрахувавши кількість інтервалів, наприклад, за формулою Стреджеса $n = 1 + [\log_2 N]$, де [.] – ціла частина числа, N – об'єм вибірки.

Позначимо інтервали через E_k, $k = \overline{1,d}$ причому, $\bigcup_{k=1}^{d} E_k = \mathbb{R}$ та позначимо через $p_k = P(E_k)$ – ймовірність потрапляння в інтервал E_k, за

умови справдження гіпотези H_0 (тобто розподіл $N(0,1)$), ν_k – кількість елементів вибірки, що потрапили в інтервал E_k, $k = \overline{1,d}$.

Складемо наступну статистику

$$\chi^2 = \sum_{k=1}^{d} \frac{(\nu_k - Np_k)^2}{Np_k}.$$

Відповідно до критерію Пірсона, ми відхиляємо гіпотезу про нормальність, якщо $\chi^2 > \chi^2_{d-1,1-\alpha}$, де $\chi^2_{d-1,1-\alpha}$ – критичне значення розподілу $\chi^2_{d-1,}$ на рівні значущості α. Зауваження: для коректного застосування критерію Пірсона слід впевнитись, що $Np_k \geq 5$.

Статистика ***тесту Шапіро-Вілка*** визначається наступним чином

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

де $x_{(i)}$ – упорядковані значення вибірки, $\bar{x}$ – вибіркове середнє, а a_i – константи, отримані з очікуваних значень та дисперсій порядкової статистики нормально розподіленої вибірки.

Значення W близьке до 1 свідчить про нормальність, менші ж значення вказують на відхилення нульової гіпотези H_0. Значення p-value нижче обраного рівня значущості (зазвичай 0.05) призводить до відхилення гіпотези H_0.

Хоча тест Шапіро-Вілка менш потужний ніж тест Колмогорова-Смірнова, він залишається корисним для візуалізації ступеня відхилення від еталонного розподілу, особливо коли підкріплюється графічним аналізом, таким як Q-Q графіки.

Розділ 3. Методологія оцінювання якості тестів за IRT

3.1 Припущення IRT

Сучасна теорія тестування базується на припущенні, що ймовірність правильної відповіді на певне завдання описується математичною функцією, яка залежить від параметрів як самої особи, так і завдання.

Припущення сучасної теорії тестування наступні:

1) Монотонність: ймовірність успішного складання завдання студентом зростає зі зростанням здібностей студента.
2) Безвимірність: модель припускає, що вимірюється одна домінантна латентна здібність, і що ця ознака є рушійною силою для відповідей, які спостерігаються для кожного завдання тесту.
3) Локальна незалежність: відповіді, дані на окремі пункти тесту, є взаємно незалежними за певного рівня здібностей.
4) Інваріантність: рівень здібностей особи не залежить від того, які саме завдання тесту їй було запропоновано, а також від конкретної вибірки осіб (з точністю до лінійного перетворення). Це дає змогу поєднувати тести, що вимірюють одну й ту саму характеристику, і порівнювати результати різних осіб, навіть якщо вони відповідали на різні завдання.
5) Якісно однорідні вибірки: ті самі функції відгуку завдань (Item Response Function, IRF) застосовуються до всіх членів вибірки.

Якщо всі припущення дотримуються, відмінності у спостереженні правильних відповідей між респондентами будуть зумовлені варіацією їхніх здібностей.

Моделі IRT прогнозують відповіді респондентів на завдання тесту на основі їхнього положення в континуумі здібностей та певних характеристик завдань, також відомих як параметри. Функція відгуку завдань характеризує цей зв'язок. Основне припущення полягає в тому, що кожна відповідь на завдання тесту дає певну оцінку рівня здібності людини. Здібності людини (далі позначатимемо через θ) простими словами – це ймовірність надання

даною особою правильної відповіді на це завдання тесту. Таким чином, чим вищі здібності людини, тим вища ймовірність правильної відповіді. Цей зв'язок можна зобразити графічно, і він відомий як крива характеристик завдання (Item Characteristic Curve, ICC).

3.2 Параметри завдання IRT

Оскільки здібності людей варіюються, їхнє положення в континуумі латентного конструкту змінюється і визначається вибіркою респондентів та параметрами завдання. Завдання має бути достатньо чутливим, щоб оцінити респондентів у межах запропонованого неспостережуваного континууму.

Складність завдання (b_i) – це параметр, який визначає поведінку завдання на шкалі здібностей. Вона визначається в точці медіанної ймовірності, тобто на рівні, за якого 50% респондентів надають правильну відповідь. На кривій характеристик завдання складні завдання зміщуються праворуч шкали, тим самим вказуючи вищі здібності респондентів, які відповіли правильно на це завдання. Разом з тим легші завдання зміщуються ліворуч шкали здібностей.

Дискримінативність завдання (a_i) визначає швидкість, з якою змінюється ймовірність надання студентом правильної відповіді залежно від його рівня здібностей. Цей параметр є важливим для диференціації між особами, які мають подібні рівні латентного конструкту, що досліджується. Кінцевою метою розробки якісного тесту є включення завдань з високою розрізнювальною здатністю, для того щоб мати змогу відобразити осіб вздовж континууму латентної ознаки. Також важливо зазначити, що дослідникам слід бути обережними, якщо спостерігається від’ємна дискримінативність завдання, оскільки ймовірність надання правильної відповіді не повинна зменшуватися зі збільшенням здібностей респондента.

Вгадування завдання (c_i) – це третій параметр, який враховує ймовірність вгадування правильної відповіді на завдання. Він обмежує

ймовірність надання правильної відповіді, коли здібність наближається до $-\infty$.

Враховуючи одне з припущень, а саме інваріантність популяції: параметри завдання поводяться однаково на різних популяціях. Варто зазначити, що це не справджується в класичній теорії тестування, проте оскільки в IRT одиницею аналізу є саме завдання, а не цілий тест, складність завдання може бути стандартизована (пройти лінійне перетворення) в різних популяціях, і таким чином завдання можна буде легко порівнювати. Важливо додати, що навіть після лінійного перетворення оцінки параметрів отримані з двох різних вибірок не будуть ідентичними: інваріантність, як випливає з назви, стосується інваріантності популяції, тому вона застосовується лише до параметрів популяції завдання.

3.3 Типи моделей IRT

Однопараметрична логістична модель є найпростішою формою моделі в IRT. Вона складається лише з одного параметра, що описує латентну рису (здібність – θ) людини, яка відповідає на завдання, а також іншого параметра для завдання – складність.

Наступне рівняння представляє її математичну форму

$$P(Y_{is} = 1|\theta_s) = \frac{exp\{a \cdot (\theta_s - b_i)\}}{1 + exp\{a \cdot (\theta_s - b_i)\}},$$

де θ_s – здібність $s-$го студента, b_i – складність $i-$го завдання, a – параметр дискримінативності (в цьому випадку це константа – всі завдання мають однаковий рівень складності).

Ця модель представляє функцію відгуку завдання для однопараметричної логістичної моделі, яка прогнозує ймовірність правильної відповіді залежно від здібностей респондента та складності запитання. У моделі 1-PL параметр дискримінативності фіксований для всіх запитань і отже всі криві характеристик запитань, що відповідають різним завданням у тесті, паралельні вздовж шкали здібностей.

Функція інформації тесту (Test Information Function, TIF) – це сума ймовірностей надання правильних відповідей на всі завдання тесту, відповідно ця характеристика оцінює очікуваний результат тесту.

Функція інформації завдання (Item Information Function, IIF) показує обсяг інформації, який надає кожне завдання, і розраховується шляхом множення ймовірності надання правильної відповіді на ймовірність неправильної відповіді

$$I_i(\theta; b_i) = P_i(\theta; b_i) \cdot Q_i(\theta; b_i)\ ,$$

де θ – здібність студента, b_i – параметр складності $i-$го завдання, P_i – функція ймовірності надання студентом з рівнем здібностей θ правильної відповіді на завдання з рівнем складності b_i; Q_i – функція ймовірності надання студентом з рівнем здібностей θ неправильної відповіді на завдання з рівнем складності b_i.

Слід зазначити, що кількість інформації на заданому рівні здібностей є оберненою величиною до її дисперсії, отже, чим більша кількість інформації надається завданням, тим більша точність вимірювання. Задачі, виміряні з більшою точністю, надають більше інформації та графічно зображені довшими та вужчими порівняно з їхніми аналогами, які надають менше інформації. Вершина кривої відповідає значенню b_i – здібності в точці медіани ймовірності. Максимальна кількість запропонованої інформації буде надана, коли ймовірність правильної та неправильної відповіді буде рівною, тобто 50%. Задачі є найбільш інформативними серед респондентів, які представляють весь латентний континуум, і особливо серед тих, хто має 50% ймовірність відповісти будь-яким із варіантів.

Функції інформації завдань зазвичай мають форму дзвона. Високорозрізнювальні завдання мають високі вузькі функції інформації: вони роблять великий внесок, але у вузькому діапазоні. Менш розрізнювальні завдання надають менше інформації, але над ширшим діапазоном. Графіки функцій інформації завдань можна використовувати для оцінювання того,

скільки інформації завдання вносить, й у якому діапазоні шкали оцінок. Завдяки локальній незалежності, функції інформації завдань адитивні. Відтак, функція інформації тесту є просто сумою функцій інформації завдань на іспиті.

Оцінка здібностей. Припущення локальної незалежності стверджує, що відповіді на запитання повинні бути незалежними та пов'язаними лише здібностями. Це дозволяє нам оцінити функцію правдоподібності індивідуального шаблону відповідей для тесту, що використовується, шляхом множення ймовірностей відповідей на запитання. Далі, за допомогою ітеративного процесу, обчислюється оцінка максимальної правдоподібності здібностей. Простіше кажучи, оцінка максимальної правдоподібності надає нам очікувані бали для кожної людини.

Модель Раша та логістичні моделі з одним параметром є математично рівноцінними, проте модель Раша обмежує коефіцієнт дискримінативності завдань (a_i) 1, тоді як логістична модель з одним параметром прагне максимально підібрати дані та не обмежує коефіцієнт дискримінативності. Модель Раша є кращою, оскільки вона більше зосереджується на розробці змінної, що використовується для вимірювання досліджуваного конструкту.

Наступне рівняння представляє її математичну форму

$$P_{ij}(\theta_j,\ b_i) = \frac{exp\{\theta_j - \ b_i\}}{1 + exp\{\theta_j - \ b_i\}},$$

де θ_j – здібність j−го студента, b_i – складність i − го завдання.

Двопараметрична логістична модель прогнозує ймовірність успішної відповіді використовуючи два параметри (складність b_i та дискримінативність a_i). Параметр дискримінативності може змінюватися між завданнями. Отже, криві характеристик завдання різних завдань можуть перетинатися та мати різні кути нахилу. Чим крутіший нахил, тим вища дискримінативність завдання, оскільки воно зможе виявляти незначні відмінності у здібностях респондентів.

Наступне рівняння представляє математичну форму двопараметричної моделі

$$P(Y_{is} = 1|\theta_s) = \frac{exp\{a_i \cdot (\theta_s - b_i)\}}{1 + exp\{a_i \cdot (\theta_s - b_i)\}},$$

де θ_s – здібність $s-$ го студента, b_i – складність $i-$ го завдання, a_i – параметр дискримінативності $i-$ го завдання.

Як і у випадку з моделлю 1-PL, інформація розраховується як добуток ймовірностей правильної та неправильної відповіді. Однак, тут добуток ще множиться на квадрат параметра дискримінативності. Звідси випливає, що чим більший параметр дискримінативності, тим більше інформації надає завдання. Оскільки коефіцієнт дискримінативності може змінюватися між завданнями, графіки функції інформації завдання також можуть виглядати по-різному.

$$I_i(\theta; b_i; a_i) = a_i^2 \cdot P_i(\theta; b_i; a_i) \cdot Q_i(\theta; b_i; a_i)\ ,$$

де θ – здібність студента, b_i – параметр складності $i-$ го завдання, a_i – параметр дискримінативності $i-$ го завдання, P_i – функція ймовірності надання студентом з рівнем здібностей θ правильної відповіді на завдання з рівнем складності b_i та рівнем дискримінативності a_i; Q_i – функція ймовірності надання студентом з рівнем здібностей θ неправильної відповіді на завдання з рівнем складності b_i та рівнем дискримінативності a_i.

У моделі 2-PL припущення локальної незалежності залишається в силі та використовується оцінка максимальної правдоподібності здібності. Хоча ймовірності для шаблонів відповідей все ще підсумовуються, тепер вони ще зважуються коефіцієнтом дискримінативності завдання для кожної відповіді. Тому їхні функції правдоподібності можуть відрізнятися одна від одної та досягати максимуму на різних рівнях θ.

Трипараметрична логістична модель прогнозує ймовірність правильної відповіді так само як і моделі 1-PL та 2-PL, але вона доповнена третім параметром, який називається параметром вгадування (також відомим як параметр псевдошансу), що обмежує ймовірність надання правильної

відповіді, коли здібності респондента наближаються до $-\infty$. Коли респонденти відповідають на запитання шляхом вгадування, кількість інформації, що надається цим завданням, зменшується, а функція інформації запитання досягає максимуму на нижчому рівні порівняно з іншими функціями. Крім того, складність більше не розмежована на рівні медіанної ймовірності. Запитання, на які дається відповідь шляхом вгадування, вказують на те, що здібність респондента менша за його складність.

Наступне рівняння представляє математичну форму 3-параметричної логістичної моделі

$$P(Y_{is} = 1|\theta_s) = c_i + (1 - c_i)\frac{exp\{a_i(\theta_s - b_i)\}}{1 + exp\{a_i(\theta_s - b_i)\}},$$

де θ_s – здібність $s-$го студента, b_i – складність $i-$го завдання, a_i – параметр дискримінативності $i-$го завдання, c_i – параметр вгадуваності $i-$го завдання.

Функція інформації завдання для трипараметричної моделі має вигляд

$$I_i(\theta; b_i; a_i; c_i) = a_i^2 \cdot \frac{(P_i(\theta; b_i; a_i; c_i) - c_i)^2 \cdot Q_i(\theta; b_i; a_i; c_i)}{(1 - c_i)^2 \cdot P_i(\theta; b_i; a_i; c_i)}$$

На Рис. 3.3.1 наведено типову криву характеристик для трипараметричної моделі. Вісь x показує латентну здібність (θ) у діапазоні від -4 до 4, де 0 – середня здатність у досліджуваній популяції. Вісь y показує ймовірність правильної відповіді на завдання. Таким чином, крива представляє ймовірність надання студентом з певним рівнем здібностей правильної відповіді на це завдання. Очевидно за формулюванням, що графік обмежений 1 і 0 по осі ординат. У цьому прикладі модель забезпечує плавну криву для даного завдання, оскільки вона базується на дискримінативності завдань (крутизні нахилу), складності (точка перегину, в якій ймовірність правильної відповіді на завдання становить 50%) та ймовірності вгадування (трохи підвищена нижня асимптота).

Отже, параметри моделі інтерпретують наступним чином: b – складність завдання, $p(b) = \frac{1+c}{2}$ – точка, де нахил графіку максимальний; a – дискримінативність завдання, $p'(b) = \frac{a \cdot (1-c)}{4}$ – максимальний нахил графіку; c – псевдовідгадуваність, $p(-\infty) = c$ – асимптотичний мінімум.

Тобто на графіку, чим крутіший нахил, тим краще завдання розрізняє людей, які знаходяться поблизу його рівня складності. Чим нижче розташована нижня асимптота, тим менша ймовірність випадкового вибору правильної відповіді. Оскільки складність завдання (точка перегину) трохи вища за середній рівень здібностей учасників тесту, завдання може бути розв'язане більшістю потенційних учасників.

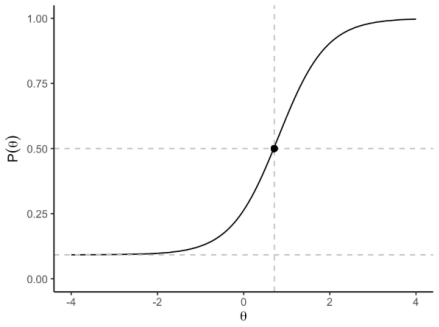


Рис. 3.3.1 Приклад кривої характеристик завдання

Відповідність моделі. Один із способів вибору моделі – це оцінити відносну відповідність моделі за допомогою її інформаційних критеріїв. Оцінки AIC порівнюються і вибирається модель з нижчим значенням.

Наступні базові джерела детальніше описують фундаментальні принципи IRT, проте вони не включають останні оновлення в теорії [15;16;17].

Варто відзначити, що питання саме сучасних підходів до теорії тестів ґрунтовно висвітлено у навчально-методичному посібнику Т. В. Лісової «Моделі та методи сучасної теорії тестів» [36]. Авторка подає матеріал глибоко та системно, тому зацікавленим у розширенні теоретичних знань (вже з врахуванням останніх оновлень) рекомендую звернутися до цієї праці.

Також К. ДеМарс [18] вдалося написати стисле, але надзвичайно інформативне видання, яке в зрозумілій манері роз'яснює найскладніші концепції IRT. Базовий посібник, що охоплює припущення, параметри та

вимоги IRT, а також пояснює, як правильно описувати отримані результати у звітах і як дослідникам слід враховувати контекст проведення тестування, популяцію респондентів та ефективне використання отриманих балів.

Ґрунтовне пояснення одновимірних та особливо багатовимірних моделей, а також їхні припущення та методологію можна прочитати у [19].

Детальніше про застосування IRT на практиці описується у роботах [21-27]. Ну і зокрема особливості застосування методології IRT в програмному середовищі R, яке я використовуватиму для опрацювання зібраних даних наведені в джерелах [20;28;29].

Розділ 4. Комплексний аналіз якості покрокових тестів з математики університетського рівня

4.1 Короткий опис досліджуваного покрокового тесту

У цьому розділі представлено приклад застосування описаної методології оцінювання якості тестів до аналізу якості стандартного покрокового студентського тесту з математичного аналізу. Головною метою цього розділу є оцінка ефективності та надійності даного формату тестування, а також визначення його особливостей. Для обробки результатів тестування використовувалося спеціальне програмне забезпечення, таке як MS Excel та R.

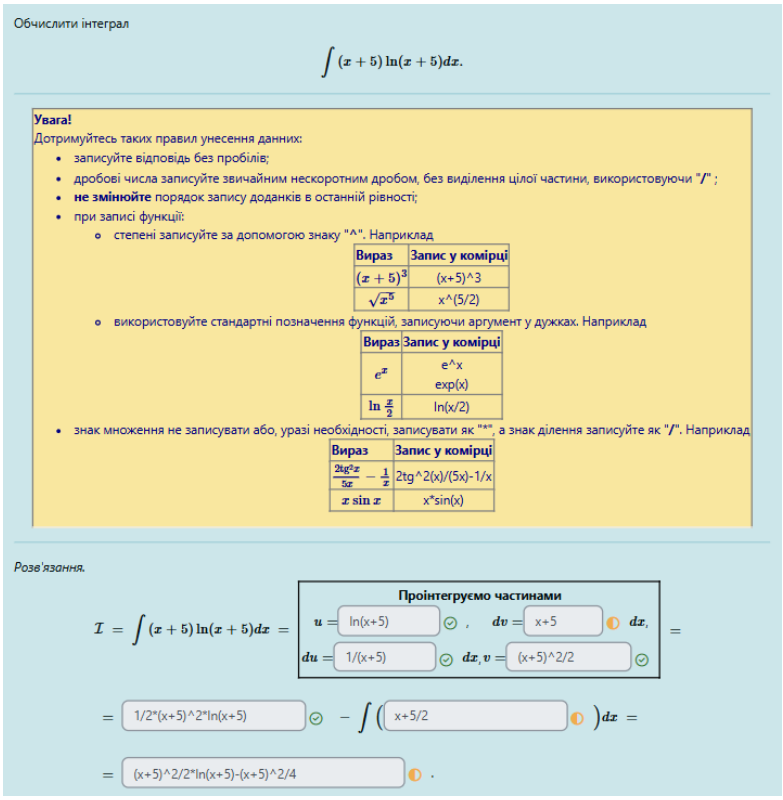


Рис. 4.1.1 Вигляд стандартного покрокового тесту

Спочатку пропоную проаналізувати структурні та практичні аспекти даного тесту. Отже, для аналізу був обраний підсумковий тест на тему «Основні методи інтегрування», що проводився для студентів першого курсу 2024/2025 н.р. за допомогою платформи Moodle. Загалом 83 студенти з факультету інформатики та обчислювальної техніки пройшли цей тест (слід зазначити, що студенти даного факультету характеризуються високим рівнем

академічної підготовки). Загалом тест складається з семи комірок, кожна з яких оцінюється в максимум один бал (Рис. 4.1.1).

Для цього тесту спочатку було створено сам шаблон, а вже потім на його основі вручну було побудовано 9 варіантів завдання. Завдяки цьому кожен варіант завдання мав однакову структуру, але різні початкові інтеграли виду $\int (x + A) \ln(x + A) dx$, де A – ціле число. У таблиці 4.1.1 наведено кількість можливих варіантів відповідей на кожен з пунктів тесту:

Частина тесту	**u**	**dv**	**du**	**v**	**uv**	**vdu**	**int**
Початкові можливі відповіді	1	2	1	2	3	2	4
Можливі відповіді після коригувань	2	2	3	6	15	6	37

Табл. 4.1.1 Кількість можливих варіантів відповідей

У стандартному тесті перевірка відповіді базується на точних збігах числових або текстових значень і бали за кожну комірку враховуються незалежно. Це вимагає ручного додавання кожного прийнятного варіанту правильної відповіді для кожної комірки, а також оформлення об’ємної пам’ятки з правилами внесення відповідей. В таблиці можна побачити кількість варіантів можливих відповідей, які є на даний час, більшість з яких було додано за час багаторічного використання даного тесту.

Крім того, у наборі даних, що використовувався для аналізу, було внесено 18 ручних виправлень балів за завдання для стандартного тесту через помилки друку. Було 41 відповідь, яку не вдалося оцінити через неправильний синтаксис введених даних, хоча в іншому вони були правильними.

4.2 Аналіз якості стандартного покрокового тесту (CTT)

На першому кроці всі результати студентів було зібрано в загальну таблицю, фрагмент якої представлено в таблиці 4.2.1. Таблицю було відсортовано за загальним балом. Максимальна оцінка для кожного завдання тесту становить один бал, загальний бал тесту нормувався шляхом поділу суми отриманих балів за кожне завдання на максимальний можливий бал за весь

тест, що забезпечує приведення результатів до єдиної шкали [0; 1]. Вважається, що кожен пункт має однакову вагу.

Students	u	dv	du	v	uv	vdu	int	Score
Student 01	1	1	1	1	1	1	1	**1.000**
Student 02	1	1	1	1	1	1	1	**1.000**
Student 03	1	1	1	1	1	1	1	**1.000**
…	…	…	…	…	…	…	…	**…**
Student 81	1	0.75	0.75	1	0	0	0	**0.500**
Student 82	1	0.75	0.75	1	0	0	0	**0.500**
Student 83	0	0.75	0.75	1	0	0	0	**0.357**

Табл. 4.2.1 Результати студентів

На наступному кроці було побудовано таблицю частот загальних балів (Табл. 4.2.2), а вже далі на її основі виконано графічне представлення даних, зокрема побудовано гістограму загальних балів та діаграму скриня-вуса (Рис 4.2.1):

Інтервали загальної оцінки	Частота
[0 ; 0.1429]	2
(0.1429 ; 0.2857]	1
(0.2857 ; 0.4286]	9
(0.4286 ; 0.5714]	2
(0.5714 ; 0.7143]	10
(0.7143 ; 0.8571]	27
(0.8571 ; 1]	32

Табл. 4.2.2 Частоти загальних оцінок

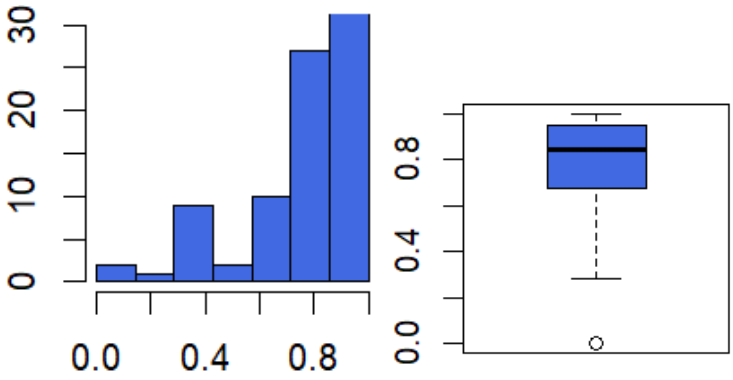


Рис. 4.2.1 Гістограма та діаграма скриня-вуса загальних оцінок

З гістограми видно, що більшість студентів отримали дуже високі бали: 50% студентів знаходяться між квартилями $Q_1 = 0.68$ та $Q_3 = 0.95$, що свідчить про високий рівень підготовки студентів. Діаграма скриня-вуса також виявляє потенційні викиди, які можуть представляти студентів, які були менш підготовлені. Однак результати цих студентів не слід виключати з аналізу.

Після побудови гістограми частот потрібно визначити міри центральних тенденцій, тобто моду, медіану та середнє вибіркове значення. Також для оцінки розподілу загальних оцінок знайдемо наступні описові статистики:

стандартне відхилення, коефіцієнт асиметрії та коефіцієнт ексцесу (Табл. 4.2.3).

Вибіркове середнє	**Медіана**	**Мода**	**Стандартне відхилення**	**Коефіцієнт асиметрії**	**Коефіцієнт ексцесу**
0.7724	0.8429	0.9600	0.2273	-1.4537	1.7973

Табл. 4.2.3 Описові характеристики загального тестового балу

Однією з ознак збалансованого тесту може бути унімодальний розподіл результатів, а також відносна близькість моди та вибіркового середнього. З таблиці 4.2.3 видно, що середнє значення (0.77) та медіана (0.84) є відносно близькими одне до одного, тоді як мода (0.96) зміщена праворуч. Отже, унімодальний розподіл результатів може свідчити про загальну збалансованість структури тесту. Коефіцієнт асиметрії (-1.45) вказує на розподіл з лівою асиметрією, а з врахуванням значення медіани (0.84) можна сказати, що більша частина студентів отримала дуже високі оцінки. Ексцес (1.80) свідчить про лептокуртичний розподіл результатів тесту, тобто вони мають гострішу вершину та важчі хвости в порівнянні з нормальним розподілом. Це означає, що дані є більш концентровані навколо середнього значення та існує більше викидів або дуже віддалених від середнього значень, ніж очікувалося б для нормального розподілу. У поєднанні з помірним стандартним відхиленням (0.23) можна зробити висновок, що студенти мають доволі високий рівень підготовки, тому бали концентруються ближче до верхньої межі.

Отже, враховуючи отримані на попередньому етапі результати вже можна очікувати відхилення гіпотези про нормальність розподілу. Власне це і підтверджується тестами Колмогорова-Смірнова ($p - value = 0.0032$) та Шапіро-Вілка ($p - value = 3.5809 \cdot 10^{-8}$). Якщо значення p-value більше за рівень значущості, який в більшості випадків вважають рівним 0.05, то в нас немає підстав прийняти нульову гіпотезу, яка припускає нормальний розподіл сумарних балів за тест.

Для розгорнутого аналізу якості тесту потрібно також окремо знайти всі описові статистики для завдань.

Комірка	u	dv	du	v	uv	vdu	int
Вибіркове середнє	0.9639	0.8855	0.7651	0.8253	0.6928	0.6235	0.6506
Стандартне відхилення	0.1878	0.2142	0.3058	0.3780	0.4609	0.4046	0.4094
Індекс легкості	0.9639	0.8855	0.7651	0.8253	0.6928	0.6235	0.6506
Індекс дискримінативності	0.0870	0.1739	0.3478	0.5435	0.9348	0.7935	0.6630
Кореляція завдання-тест	0.4416	0.5613	0.4531	0.7850	0.7599	0.7652	0.7159
Кореляція завдання-решта	0.3392	0.4582	0.2809	0.6622	0.5861	0.6216	0.5490

Табл. 4.2.4 Описові характеристики завдань тесту

Як видно з таблиці 4.2.4, середні бали для всіх завдань є вищими за 0.62, причому перші чотири пункти в цьому плані проявляють себе найкраще. Але для завдань 1 та 2 індекси дискримінативності нижчі за допустиме значення, що може свідчити про обмежену здатність цих завдань розрізняти студентів з високим та низьким рівнем знань, всі інші завдання мають високий або помірний рівень дискримінативності. Якщо дивитися з точки зору легкості завдання, то перші чотири завдання є дуже легкими (причому для першого індекс рівний 0.96, що свідчить про те, що очікувано 96% всіх студентів не залежно від рівня їхніх знань дадуть правильну відповідь на нього), отже для якіснішого оцінювання студентів їх варто допрацювати. При значенні кореляції завдання-тест більше за 0.7, завдання має високу роздільну здатність: знову ж таки присутні проблеми для завдань 1, 2 та 3.

Отже, завдання 1 та 2 мають високий індекс легкості, низький рівень дискримінативності та кореляції завдання-тест, що свідчить про те, що ці завдання є дуже легкими та не розрізняють студентів, а отже можуть потребувати доопрацювання з метою підвищення їхньої дискримінативності. Завдання 3 та 4 мають доволі високий індекс легкості та дискримінативності, а також низький рівень кореляції завдання-тест, що вказує на те, що ці завдання не дивлячись на високі середні бали мають прийнятний рівень дискримінативності, проте їх все ж бажано переглянути. Завдання 5, 6 та 7

мають відносно низький рівень легкості, високий рівень дискримінативності та кореляції завдання-тест, що визначає ці завдання найскладнішими та найінформативнішими частинами тесту. Ці пункти, ймовірно, найбільше сприяли диференціації рівнів підготовки студентів.

Ще одним не менш важливим етапом дослідження якості тестів є знаходження кореляційної матриці завдань (Рис. 4.2.2). Одразу бачимо, що всі значення елементів в матриці додатні, це свідчить про те, що завдання не суперечать одне одному. Перші чотири пункти тесту демонструють дуже слабку ($r < 0.2$) або слабку ($0.2 \leq r < 0.4$) кореляцію з рештою трьома, що в поєднанні з високими середніми балами та низькою дискримінативністю може вказувати на доцільність перегляду ролі цих пунктів у структурі тесту. Натомість останні три пункти демонструють слабку або помірну кореляцію, нижчі середні бали та вищі значення дискримінативності, що підтримує їх збереження в тесті.

Комірка	u	dv	du	v	uv	vdu	int
u	1.00	0.53	0.38	0.25	0.15	0.14	0.15
dv	0.53	1.00	0.50	0.34	0.25	0.25	0.19
du	0.38	0.50	1.00	0.09	0.11	0.21	0.16
v	0.25	0.34	0.09	1.00	0.55	0.57	0.59
uv	0.15	0.25	0.11	0.55	1.00	0.59	0.46
vdu	0.14	0.25	0.21	0.57	0.59	1.00	0.44
int	0.15	0.19	0.16	0.59	0.46	0.44	1.00

Рис. 4.2.2 Матриця кореляцій завдання-завдання

Також варто зазначити, що коефіцієнт середньої кореляції між завданнями рівний 0.3292, що може свідчити про достатній рівень збалансованості та внутрішньої узгодженості тесту.

Коефіцієнти надійності даного тесту α Кронбаха (0.7699) та ω Макдональда (0.7814) вказують на помірну надійність, що свідчить про необхідність подальшого вдосконалення тесту.

4.3 Аналіз якості стандартного покрокового тесту (IRT)

Далі пропоную розглянути аналіз якості стандартного покрокового тесту за методологією сучасної теорії тестування. Всі розрахунки

виконувалися в R з використанням спеціальних вбудованих функцій та бібліотек: основні з них це tidyverse (для обробки та візуалізації даних), mirt (для основних аналізів IRT), а також пакет ggmirt, який є розширенням mirt і містить функції для побудови більшої кількості характеристик.

Також одразу слід зазначити, що виходячи з неперервного розподілу оцінок за кожне із завдань досліджуваного нами тесту, необхідно використовувати градуйований тип моделі відгуку (graded response model). За структурою вона ідентична двопараметричній моделі (не містить параметр вгадування), проте має дещо складніші методи розрахунку. Тобто ми будуватимемо градуйовану двопараметричну модель, що враховуватиме дискримінативність (параметр a) та складність кожного із завдань (параметр b).

Щоб підібрати будь-яку модель для IRT в R, необхідно просто використати функцію mirt() та вказати в параметрах цієї функції: набір даних, що містить лише оцінки за окремі завдання для кожного студента; початкову модель (за замовчуванням = 1); тип моделі, що будується (в нашому випадку «graded»); зазначити, що потрібно/не потрібно будувати інформацію по кожній з ітерацій. Створений об'єкт має клас «SingleGroupClass» і містить всю необхідну інформацію для оцінки побудованої моделі. Запускаючи сам об'єкт ми отримуємо інформацію про тип оцінки, а також деякі індекси відповідності моделі (включаючи AIC та BIC), які можна використовувати для порівняння моделей одна з одною (Рис. 4.3.1).

```
Call:
mirt(data = items, model = 1, itemtype = "graded", verbose = FALSE)

Full-information item factor analysis with 1 factor(s).
Converged within 1e-04 tolerance after 15 EM iterations.
mirt version: 1.45.1
M-step optimizer: BFGS
EM acceleration: Ramsay
Number of rectangular quadrature: 61
Latent density type: Gaussian

Log-likelihood = -297.5777
Estimated parameters: 14
AIC = 623.1555
BIC = 657.0192; SABIC = 612.8598
G2 (113) = 46.69, p = 1
RMSEA = 0, CFI = NaN, TLI = NaN
```

Рис. 4.3.1 Індекси відповідності побудованої двопараметричної моделі

Якщо ми використаємо функцію summary(), то отримаємо результати факторного рішення (Рис. 4.3.2), що включає факторні навантаження (factor loading, F1 – показує силу зв’язку кожного завдання з латентною здібністю, що вимірюється тестом; як правило, навантаження > 0.5 вважають суттєвим) та комунальність (communality, h2 – частка дисперсії кожного із завдання, яку пояснює латентна здібність, що вимірюється тестом, вона обчислюється як квадрат факторного навантаження; значення ближче до 1 означають, що латентна здібність добре пояснює поведінку респондента у цьому завданні).

```
         F1    h2
Item1 0.597 0.357
Item2 0.464 0.215
Item3 0.214 0.046
Item4 0.925 0.855
Item5 0.879 0.773
Item6 0.621 0.386
Item7 0.616 0.380

SS loadings:  3.012
Proportion Var:  0.43
```

Рис. 4.3.2 Результати факторного рішення

Майже всі завдання в нашому випадку мають суттєвий зв'язок з латентною здібністю (Рис 4.3.2), завдання 2 та особливо 3 мають слабший зв’язок. Завдання 3 має лише приблизно 5% поясненої дисперсії, що може вказувати на слабкий зв’язок цього завдання з основною латентною здібністю. SS loadings (3.012) – це загальна пояснена дисперсія латентною здібністю. Proportion Var (0.43) – це частка загальної дисперсії, яку пояснює латентна здібність. Тобто, модель має одну суттєву латентну здібність, що пояснює ~43% дисперсії відповідей на завдання.

Однак у IRT нас зазвичай більше цікавлять фактичні параметри завдання, такі як дискримінативність, складність та ймовірність вгадування (нагадуємо, її в нашій моделі немає).

```
$items
          a     b1
Item1 1.268 -3.172
Item2 0.891  0.168
Item3 0.374 -1.009
Item4 4.135 -0.981
Item5 3.141 -0.521
Item6 1.348  0.329
Item7 1.332  0.381
```

Рис. 4.3.3 Параметри завдань побудованої моделі

Як бачимо (Рис. 4.3.3), значення параметрів нахилу (a-параметри) коливаються від 0.374 до 4.135. Цей параметр є мірою того, наскільки добре завдання диференціює людей з різними рівнями латентної здібності. Більші значення або відповідно крутіші нахили графіку краще диференціюють людей. Найвищі значення дискримінативності мають завдання 4 та 5, далі йдуть 6, 7 та 1, завдання 2 та 3 майже не розрізняють учнів з високим та низьким рівнем здібностей. Параметр складності (b-параметр) інтерпретується як значення латентної здібності, що відповідає 50% ймовірності правильної відповіді на це запитання. Параметри складності показують, що загалом завдання охоплюють доволі широкий діапазон латентної ознаки: завдання 1 є дуже легким – на нього очікувано дадуть правильну відповідь навіть студенти з дуже низьким рівнем здібностей; завдання 3, 4 та 5 теж мають від'ємні значення, що означає відносну легкість цих завдань; завдання 6 та 7 вже потребують вищих рівнів здібностей студента, завдання 2 не дуже добре розрізняє рівень знань – значення близьке до 0.

Отже завдання 1 є дуже легким, але при цьому непогано диференціює зовсім слабких студентів: чітко відділяє тих, хто взагалі не має базового рівня здібностей, тобто його можна залишити в тесті, щоб одразу розрізняти студентів з найнижчим рівнем знань. Завдання 2 має середню складність і не дуже добре розрізняє рівень знань: бажано його скорегувати. Завдання 3 є легким і найменш дискримінативним: доцільно розглянути можливість його суттєвого перегляду або виключення із тесту. Завдання 4 і 5 є відносно легкими, проте найбільш дискримінативними: дуже добре розрізняють студентів з низьким та середнім рівнем здібностей. Завдання 6 та 7 є складними та помірно дискримінативними: непогано розрізняють студентів з середнім та високим рівнем здібностей.

Криві характеристик завдань візуалізують усі параметри IRT (у нашому випадку цих параметрів 2) для всіх завдань тесту (Рис. 4.3.4). Дана візуалізація допомагає краще зрозуміти унікальні властивості кожного завдання, а також може бути корисною для виявлення прогалин в оцінюванні.

Інший спосіб оцінити якість кожного завдання – це побудувати так звані криві інформації завдання (Рис. 4.3.5). Інформація – це статистичне поняття, яке стосується здатності завдання якісно виставляти бали за здібністю студента. Інформація на рівні завдання уточнює, наскільки добре кожен елемент тесту сприяє точності загальної оцінки, причому вищі рівні інформації призводять до точніших оцінок.

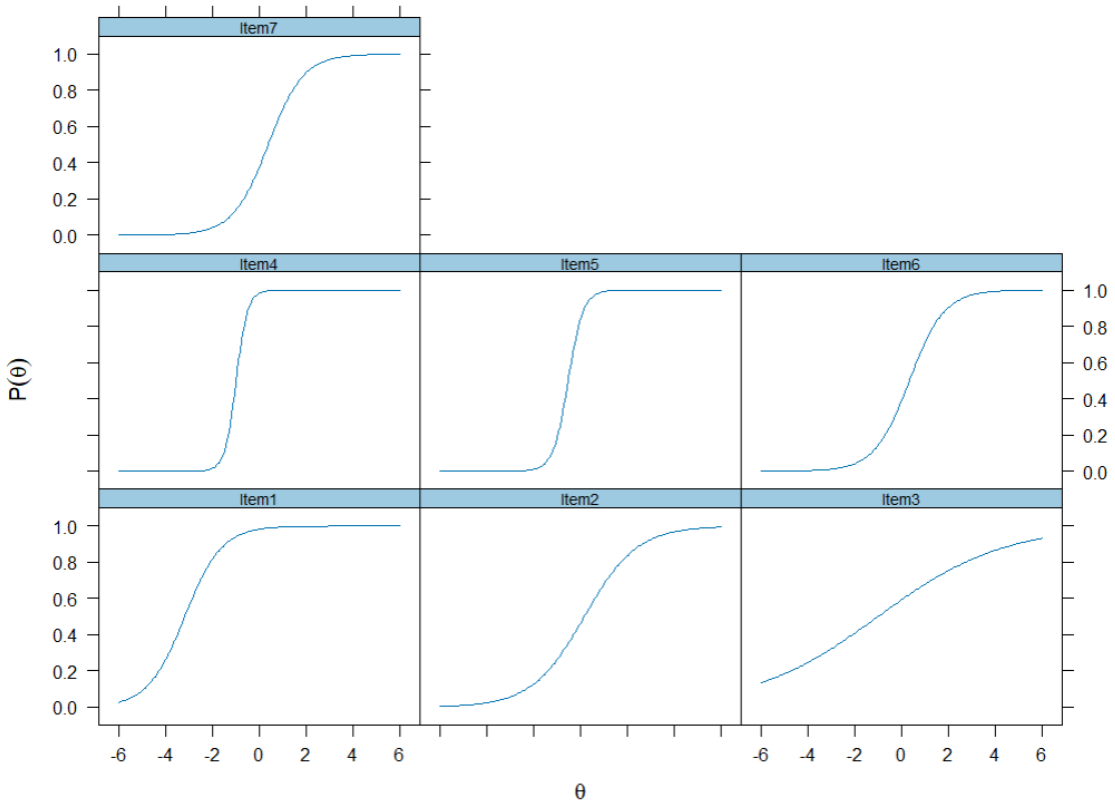


Рис. 4.3.4 Криві характеристик завдань

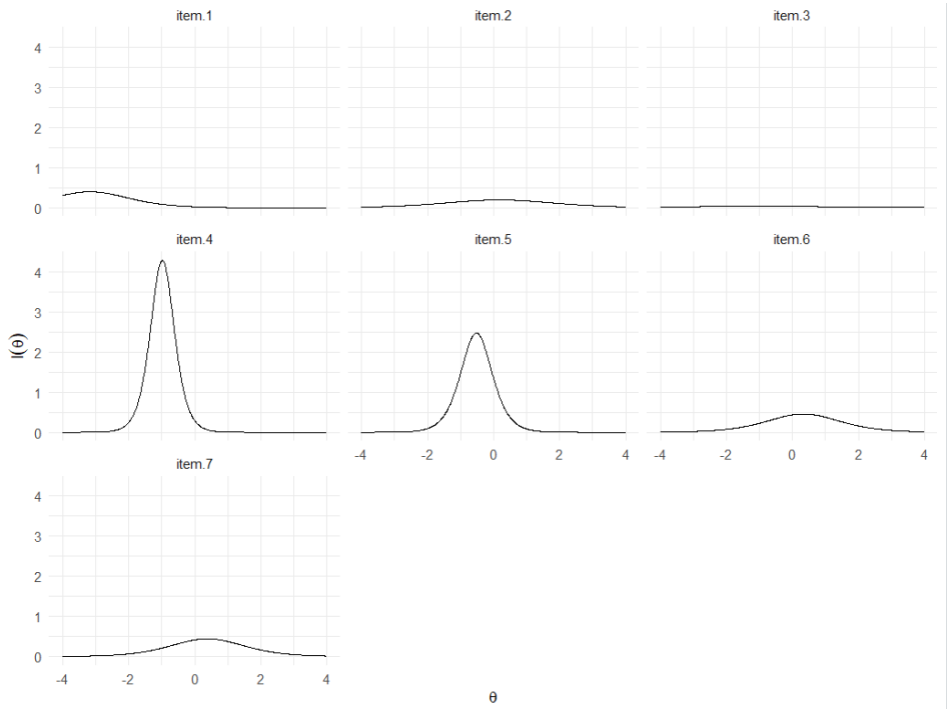


Рис. 4.3.5 Криві інформації завдань

Подібно до факторно-аналітичних підходів, ми можемо оцінити, наскільки добре модель відповідає наявним даним. Проте тут замість статистики χ^2 використовується специфічна статистика M2, спеціально розроблена для оцінки відповідності моделей відповідей на завдання.

```
> M2(mod_grm)
            M2 df         p RMSEA RMSEA_5   RMSEA_95      SRMSR      TLI CFI
stats 12.8754 14 0.5363574     0       0 0.09862786 0.07473497 1.019335   1
```

Рис. 4.3.6 Статистики відповідності побудованої моделі

Як бачимо, статистика M2 є порівняно низькою та незначущою (Рис. 4.3.6). Тому не виявлено статистично суттєвих відмінностей між побудованою моделлю та зібраними даними. Це додатково підтверджується дуже низьким значенням RMSEA, CFI та TLI рівним 1.

Однак в IRT нас зазвичай більше цікавлять індекси відповідності окремих завдань та здібностей осіб. IRT дозволяє нам оцінити, наскільки добре кожне завдання відповідає моделі та чи відповідають патерни відповідей осіб з певним рівнем здібностей моделі.

Почнемо з перевірки відповідності завдань: можемо використати функцію itemfit(), щоб отримати різноманітні показники відповідності. За

замовчуванням ми отримуємо S_X2 Орландо і Тіссена та відповідні значення dfs, RMSEA та p-value (це значення має виходити незначущим, щоб вказувати на хорошу відповідність завдань побудованій моделі – більшим за 0.05).

	item	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2
1	Item1	NA	NA	NA	NA
2	Item2	1.816	3	0.000	0.611
3	Item3	2.620	3	0.000	0.454
4	Item4	NaN	0	NaN	NaN
5	Item5	2.832	1	0.149	0.092
6	Item6	1.782	2	0.000	0.410
7	Item7	0.752	2	0.000	0.687

Рис. 4.3.7 Статистики відповідності отриманих параметрів завдання

Як ми бачимо, для першого та четвертого завдання статистика не могла бути обчислена через відсутність варіацій у відповідях – майже всі відповіді (80 із 83) були однакові та правильні. Отже, ці пункти доцільно переглянути з огляду на обмежену варіативність відповідей. Всі інші завдання демонструють високу відповідність побудованій моделі (Рис. 4.3.7).

Прихильники моделі Раша дуже часто замість статистики S_X2 використовують статистику infit та outfit. Ми можемо вивести їх, додавши до функції аргумент fit_stats=infit. У результаті отримуємо як середньоквадратичну, так і стандартизовану версії цих статистик. Infit (information-weighted fit) показує, чи певне завдання поводиться стабільно у своїй зоні здібностей θ (де він є найбільш інформативним). Outfit (outlier-sensitive fit) показує, чи є аномальні відповіді, які модель не може пояснити (цей показник чутливий до аномальних відповідей саме у віддалених діапазонах θ – студенти з низьким рівнем здібностей відповіли правильно, або навпаки студенти з високим рівнем знань дали неправильну відповідь на завдання). У пакеті ggmirt ми також можемо використовувати функцію itemfitPlot() для візуальної перевірки цього фактору.

З точки зору інтерпретації отриманих нестандартизованих результатів: > 2.0: завдання спотворює систему оцінювання (може бути спричинено кількома аномальними відповідями – «викидами»); $1.5 - 2.0$: завдання непродуктивне для побудови тесту, але не спотворює його (завдання не додає інформації, проте не шкодить моделі); $0.5 - 1.5$: оптимальний діапазон

(завдання добре узгоджується з моделлю); < 0.5: менш продуктивний для тесту, але не спотворює модель (завдання занадто передбачуване, може створювати оманливо високі показники надійності та роздільності).

	item	outfit	z.outfit	infit	z.infit
1	Item1	1.026	0.359	1.049	0.273
2	Item2	0.872	-1.471	0.907	-1.363
3	Item3	0.979	-0.376	0.981	-0.338
4	Item4	0.237	-0.869	0.603	-1.066
5	Item5	0.334	-1.172	0.714	-1.319
6	Item6	0.708	-1.943	0.810	-2.285
7	Item7	0.713	-1.876	0.811	-2.295

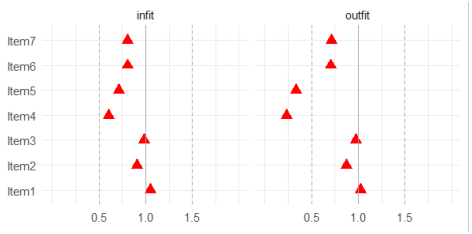
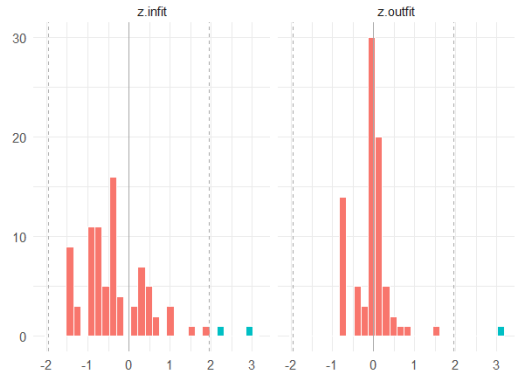


Рис. 4.3.8 Статистики (infit-outfit) відповідності отриманих параметрів

Ми бачимо, що всі завдання мають оптимальні статистки infit та outfit, це говорить про те, що вони є добре узгодженими з побудованою моделлю. Для завдань 4 та 5 є проблеми, але вони не є суттєвими (Рис. 4.3.8).

Оцінка відповідності особи. Технічно ми можемо створити точно такі ж показники для кожної особи, щоб оцінити, наскільки добре моделі відповідей кожної особи узгоджуються з загальною побудованою моделлю. Тобто ми перевіряємо, чи не має наша модель аномалій: якщо людина з високими латентними здібностями відповідає неправильно на просте запитання, ця людина не відповідає моделі; і навпаки, якщо людина з низькими здібностями правильно відповідає на дуже складне запитання, вона також не відповідає нашій моделі. На практиці, найімовірніше, буде кілька людей, які не відповідають моделі добре. Але якщо кількість респондентів, які не відповідають вимогам невелика, то це не впливає на якість побудованої моделі. Ми здебільшого знову будемо дивитися на статистики infit та outfit: якщо менше 5% респондентів мають значення по модулю більші за 1.96, то модель узгоджується зі здібностями особи.



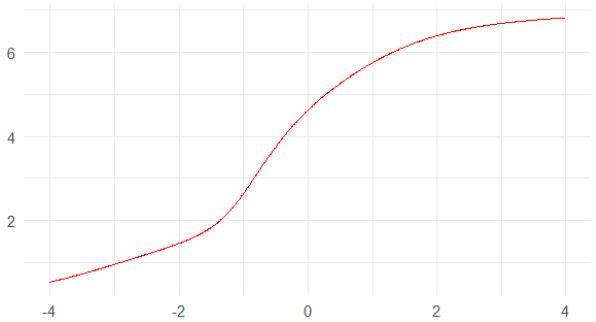

```
infit.outside outfit.outside
   0.97590361     0.98795181
   0.02409639     0.01204819
```

Рис. 4.3.9 Статистики (infit-outfit) відповідності здібності особи

Як бачимо, лише 2.4% та 1.2% для infit та outfit статистик відповідно знаходяться за межами інтервалу оптимальних значень (Рис. 4.3.9, на графіку ці викиди зображено блакитним кольором), отже більшість респондентів демонструють узгодженість відповідей із побудованою моделлю (навіть занадто узгоджено – можливо, тест є дуже простим для цих студентів).

Останньою характеристикою якості IRT-моделі може бути те, що загальний тестовий бал за побудованою моделлю для кожного студента є достатньо хорошою оцінкою здібності. Графік так званої характеристичної кривої шкали дозволяє оцінити це візуально, відображаючи зв'язок між рівнем здібностей та сумарним оціночним балом. Ця крива зазвичай має S-подібну форму, оскільки зв'язок сильніший у середньому діапазоні здібностей та гірший на крайніх його значеннях.

Звичайно, ми також можемо перевірити це за допомогою простого обчислення кореляції. Спочатку витягуємо латентний бал IRT за допомогою функції fscores(), потім співвідносимо його зі спостережуваним (Рис. 4.3.10). Як бачимо, вони мають дуже високу кореляцію (0.92)




```
        Pearson's product-moment correlation

data:  score and sumscore
t = 20.665, df = 81, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8739628 0.9455362
sample estimates:
      cor
0.9168263
```

Рис. 4.3.10 Оцінка відповідності побудованої моделі

4.4 Загальні висновки щодо якості покрокового тесту

Перші чотири пункти мало сприяють диференціації, тоді як останні три забезпечують найбільшу дискримінативність. Оскільки всі сім пунктів з'являються в розв’язуваннях подібних задач в письмових контрольних роботах, рекомендується зберегти всі пункти, але призначити ваги на основі складності. Також було б корисно зібрати більше даних від студентів з різним рівнем підготовки, щоб отримати більш репрезентативну вибірку та забезпечити більш точну та надійну оцінку якості тесту.

Проведений аналіз результатів тестування студентів дозволяє зробити низку узагальнень щодо ефективності запропонованого тестового інструментарію та особливостей його функціонування в умовах академічного середовища. Розгляд отриманих даних у контексті як класичної, так і сучасної теорії тестування свідчить, що тест загалом виконує своє основне призначення та може забезпечувати достатньо об’єктивне вимірювання навчальних досягнень студентів і водночас відображає певні закономірності, притаманні досліджуваним респондентам.

Передусім варто зазначити, що високі результати, продемонстровані більшістю студентів, не можна тлумачити виключно як свідчення недостатньої складності тесту чи недосконалості його структури, адже конкретно у даному випадку домінуючим чинником виступає високий рівень академічної підготовки студентів факультету. Це об’єктивно впливає на результати тестування, зумовлюючи високу концентрацію балів у верхньому діапазоні шкали. Такі результати свідчать не стільки про легкість тесту, скільки про збалансованість його структури щодо реального рівня підготовленості студентів. Фактично тест виявився адекватним до академічного контингенту, для якого він був розроблений. Це є позитивним показником, адже він засвідчує валідність тесту – його здатність коректно вимірювати те, що він має вимірювати, у межах певної цільової групи.

Разом з тим результати аналізу спонукають до роздумів про можливості подальшого вдосконалення тестових матеріалів. Навіть за умови високої успішності студентів доцільно прагнути до підвищення дискримінативної здатності тесту, щоб він міг якомога точніше розрізняти між собою студентів із високим, достатнім та відмінним рівнем знань. Для цього до тесту, наприклад, можна додати кілька завдань підвищеної складності – таких, що потребують не лише відтворення вивченого матеріалу, а й творчого застосування знань у нових контекстах, побудови власних логічних аргументацій або інтерпретацій математичних результатів. Це сприятиме формуванню більш розгорнутої діагностичної шкали, зокрема у верхньому діапазоні здібностей.

Ще однією важливою тенденцією, яка простежується за результатами дослідження, є висока внутрішня узгодженість тесту. Завдання тісно корелюють між собою, що свідчить про наявність спільного конструкта, який лежить в основі оцінювання. Це означає, що всі пункти спрямовані на вимірювання єдиного латентного чинника – рівня засвоєння студентами основних понять і методів математичного аналізу для обчислення інтегралів. Така структурна цілісність підтверджує правильність підбору завдань і логічність побудови тесту. У той же час деякі завдання показали дещо нижчий рівень узгодженості з іншими, що може бути пов’язано або з відмінностями у змісті, або з різницею у когнітивному навантаженні. Подальша оптимізація тесту може полягати у ретельнішому доборі таких завдань, аби забезпечити їх концептуальну однорідність.

Отже, аналіз тесту загалом, а також його окремих завдань дозволяє говорити про достатньо високу надійність тесту, що є важливою умовою при використанні його як інструменту для оцінювання навчальних досягнень. Це означає, що результати тесту є стабільними та відтворюваними, а отже, можуть служити достовірною основою для педагогічних висновків.

Розділ 5. Комплексний аналіз якості тестів на основі STACK з математики університетського рівня

5.1 Короткий опис досліджуваного тесту на основі STACK

У цьому розділі представлено приклад застосування описаної методології оцінювання якості тестів до аналізу якості студентського тесту на основі STACK з математичного аналізу. Головною метою цього розділу є оцінка ефективності та надійності даного формату тестування, а також визначення його особливостей. Знову ж таки, для обробки результатів тестування використовувалося спеціальне програмне забезпечення, таке як MS Excel та R.

Отже, для аналізу був обраний підсумковий тест на тему «Основні методи інтегрування», що проводився для студентів першого курсу 2024/2025 н.р. за допомогою платформи Moodle. Загалом 89 студенти з факультету інформатики та обчислювальної техніки пройшли цей тест (слід зазначити, що студенти даного факультету характеризуються високим рівнем академічної підготовки). Загалом тест складається з семи комірок, кожна з яких оцінюється в максимум один бал (Рис. 5.1.1).

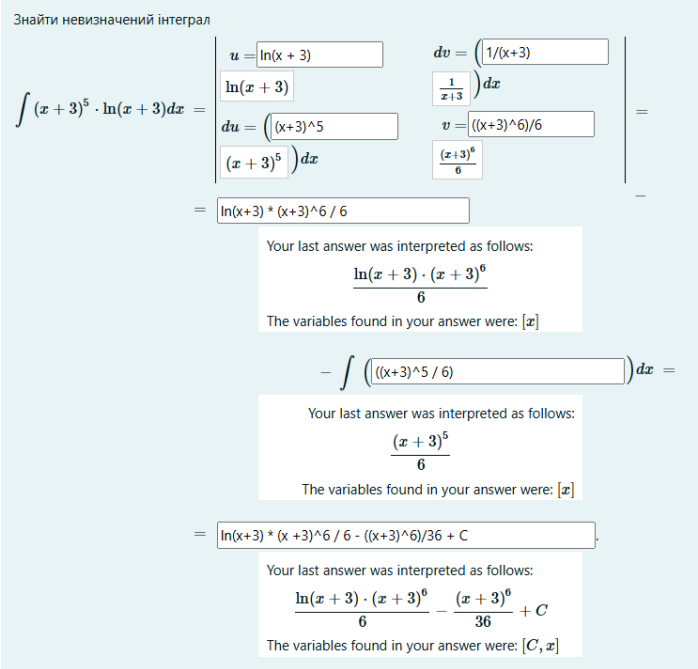


Рис. 5.1.1 Вигляд тесту на основі STACK

Також варто зазначити, що взагалі-то комірки можуть оцінюватися в залежності одна від одної за допомогою побудови дерев розв’язків. Проте для нашого аналізу ми обрали спрощений варіант тесту на основі STACK, де всі комірки оцінювалися незалежно.

Цей тест вимагав лише один початковий інтеграл виду $\int (x+A)^B \cdot \ln(x+A)dx$. Значення A та B генерувалися системою випадково для кожного варіанту тесту з наступних множин: $A \in \{2,3,\ldots,9\}$ та $B \in \{-5,-4,\ldots,4,5\}$. Це означає, що лише один шаблон на основі STACK може генерувати 704 унікальні задачі. Однак розробка тесту на основі STACK є складнішою, оскільки вимагає побудови детальних дерев рішень для обробки часткових розв’язків. У таблиці 5.1.1 наведено кількість можливих відповідей на кожен з пунктів тесту.

Частина тесту	**u**	**dv**	**du**	**v**	**uv**	**vdu**	**int**
Початкові можливі відповіді	1	1	1	1	1	2	2

Табл. 5.1.1 Кількість можливих варіантів відповідей

Як видно з таблиці, кількість можливих відповідей на кожне завдання в цьому тесті не обов'язково повинна включати різні способи представлення відповідей. Це пов'язано з символьною перевіркою, яка використовується в тесті на основі STACK. Ба більше, створюючи відповідні дерева рішень, ми можемо зробити перевірку рішень більш гнучкою до наданих відповідей. Крім того, символьна перевірка відповідей тесту на основі STACK та вбудований попередній перегляд допомагають зменшити кількість помилок друку. У наборі даних, що використовувався для аналізу, було внесено лише 2 ручні виправлення. Це також пояснює, чому в результатах тесту сталося лише 4 помилки введення.

5.2 Аналіз якості покрокового тесту на основі STACK (CTT)

Аналогічно до аналізу стандартних покрокових тестів, спочатку було зібрано таблицю результатів проходження тесту (Табл. 5.2.1) на основі якої побудовано таблицю частот загальних балів (Табл. 5.2.2).

Students	u	dv	du	v	uv	vdu	int	Score
Student 01	1	1	1	1	1	1	1	**1.000**
Student 02	1	1	1	1	1	1	1	**1.000**
Student 03	1	1	1	1	1	1	1	**1.000**
…	…	…	…	…	…	…	…	**…**
Student 87	1	0.75	0.75	1	0	0	0	**0.500**
Student 88	1	0.75	0.75	1	0	0	0	**0.500**
Student 89	0	0.75	0.75	1	0	0	0	**0.357**

Табл. 5.2.1 Результати студентів

Далі на основі таблиці частот було виконано графічне представлення даних, побудовано гістограму загальних балів, а також діаграму скриня-вуса (Рис. 5.2.1).

Інтервали загальної оцінки	**Частота**
[0 ; 0.1429]	4
(0.1429 ; 0.2857]	0
(0.2857 ; 0.4286]	0
(0.4286 ; 0.5714]	3
(0.5714 ; 0.7143]	5
(0.7143 ; 0.8571]	8
(0.8571 ; 1]	69

Табл. 5.2.2 Частоти загальних оцінок

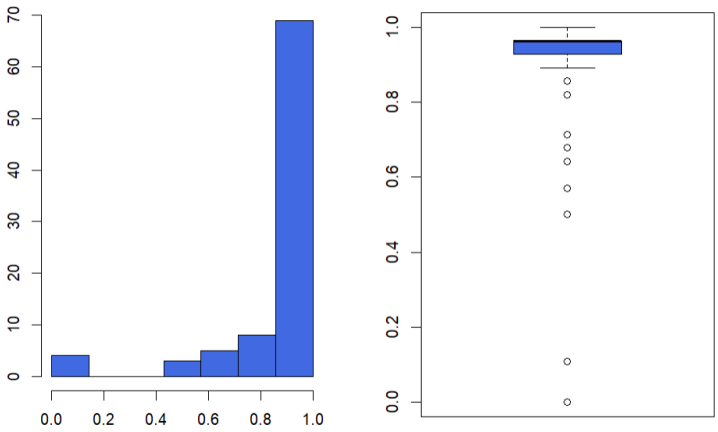


Рис. 5.2.1 Гістограма та діаграма скриня-вуса загальних оцінок

З гістограми видно, що більшість студентів отримали найвищий бал – 1, 50% студентів мають оцінку вище 0.96, що свідчить про дуже високий рівень підготовки студентів. Діаграма скриня-вуса також виявляє потенційні викиди, що можуть представляти студентів, які були менш підготовлені: як і попередній раз цих студентів з аналізу не виключаємо.

Після побудови гістограми частот знаходимо потрібні для аналізу вибіркові статистики (Табл. 5.2.3). З таблиці видно, що середнє значення (0.87), медіана (0.96) та мода (0.96) є близькими одне до одного: унімодальний розподіл результатів може свідчити про загальну збалансованість тесту. Коефіцієнт асиметрії (-3.09) вказує на розподіл з великою лівою асиметрією, а з врахуванням значення медіани (0.96) можна сказати, що майже всі студенти отримали максимальний бал. Ексцес (9.36) свідчить про лептокуртичний

розподіл результатів тесту, тобто вони мають дуже вузький і високий центр та довгі та тонкі хвости з обох боків. Це означає, що дані є більш концентровані навколо середнього значення та існує більше викидів або дуже віддалених від середнього значень, ніж очікувалося б для нормального розподілу. У поєднанні з помірним стандартним відхиленням (0.21) можна припустити, що студенти цієї вибірки мають високий рівень підготовки, а саме тому майже всі бали сконцентровані біля верхньої межі.

Вибіркове середнє	**Медіана**	**Мода**	**Стандартне відхилення**	**Коефіцієнт асиметрії**	**Коефіцієнт ексцесу**
0.8864	0.9643	0.9600	0.2146	-3.0906	9.3358

Табл. 5.2.3 Описові характеристики загального тестового балу

Отже, враховуючи отримані на попередньому етапі результати можна одразу очікувати відхилення гіпотези про нормальність розподілу. Власне це і підтверджується тестами Колмогорова-Смірнова ($p - value = 2.8628 \cdot 10^{-9}$) та Шапіро-Вілка ($p - value = 1.5489 \cdot 10^{-15}$).

Для розгорнутого аналізу якості тесту потрібно також знайти окремо всі описові статистики завдань.

Комірка	**u**	**dv**	**du**	**v**	**uv**	**vdu**	**int**
Вибіркове середнє	0.9551	0.9129	0.9073	0.9326	0.8989	0.7500	0.8483
Стандартне відхилення	0.2084	0.2641	0.2675	0.2522	0.3032	0.2820	0.3167
Індекс легкості	0.9551	0.9129	0.9073	0.9326	0.8989	0.7500	0.8483
Індекс дискримінативності	0.1600	0.3100	0.3300	0.2400	0.3600	0.4800	0.4600
Кореляція завдання-тест	0.8739	0.8404	0.8184	0.8169	0.8444	0.7259	0.6707
Кореляція завдання-решта	0.8342	0.7750	0.7443	0.7475	0.7681	0.6162	0.5270

Табл. 5.2.4 Описові характеристики завдань тесту

Як видно з таблиці 5.2.4, середні бали для всіх завдань є вищими за 0.75, відповідно всі завдання мають завищений індекс легкості. Для першого завдання індекс дискримінативності є нижчим за його нижню оптимальну межу, завдання 2, 3, 4 та 5 мають середній індекс дискримінативності, завдання 6 і 7 – високий. Кореляція завдання-тест більша за 0.7 для всіх завдань, окрім 7 (0.67) – всі вони мають високу роздільну здатність.

Отже, перше завдання має дуже високий індекс легкості й низький рівень дискримінативності, що свідчить про те, що це завдання є дуже легким та погано розрізняє студентів – може потребувати допрацювання з метою підвищення дискримінативності. Завдання 2, 3, 4 та 5 натомість мають високий індекс легкості й оптимальний рівень дискримінативності, що вказує на те, що ці завдання не дивлячись на високі середні бали мають прийнятний рівень роздільності – проте доцільно розглянути можливість їхнього перегляду. Завдання 6 та 7 мають відносно низький рівень легкості та високий рівень дискримінативності: ці завдання є найскладнішими та найінформативнішими частинами тесту і, ймовірно, найбільше сприяли диференціації рівнів підготовки студентів.

Ще одним не менш важливим етапом дослідження якості тестів є знаходження кореляційної матриці завдань (Рис. 5.2.2). Одразу бачимо, що всі значення елементів в матриці додатні, це свідчить про те, що завдання не суперечать одне одному. Усі пункти тесту мають помірну ($0{,}4 \leq r < 0{,}7$) або дуже високу ($0{,}7 < r$) взаємну кореляцію, що взагалі-то підтримує їх збереження в тесті. Проте перші чотири пункти тесту демонструють меншу кореляцію з рештою трьома, що в поєднанні з високими середніми балами та низькою дискримінативністю може вказувати на доцільність перегляду цих пунктів тесту. Натомість останні три завдання демонструють помірну кореляцію, нижчі середні бали та вищі значення дискримінативності, що точно підтримує їх збереження в тесті.

Комірка	u	dv	du	v	uv	vdu	int
u	1.00	0.75	0.74	0.81	0.65	0.58	0.46
dv	0.75	1.00	0.74	0.76	0.67	0.37	0.45
du	0.74	0.74	1.00	0.58	0.72	0.50	0.34
v	0.81	0.76	0.58	1.00	0.65	0.48	0.37
uv	0.65	0.67	0.72	0.65	1.00	0.53	0.46
vdu	0.58	0.37	0.50	0.48	0.53	1.00	0.55
int	0.46	0.45	0.34	0.37	0.46	0.55	1.00

Рис. 5.2.2 Матриця кореляцій завдання-завдання

Також варто зазначити, що коефіцієнт середньої кореляції між завданнями рівний 0.5789, що свідчить про можливе дублювання питань – надто схожі питання потрібно або прибрати з тесту, або скорегувати.

Коефіцієнти надійності даного тесту α Кронбаха (0.8978) та ω Макдональда (0.9095) вказують на добру надійність тесту, проте враховуючи показники по завданням окремо, тест може потребувати перегляду та певної модифікації з урахуванням характеристик окремих завдань.

5.3 Аналіз якості покрокового тесту на основі STACK (IRT)

Повторюємо всі кроки аналогічно до аналізу стандартного покрокового тесту. Спочатку виводимо результати факторного рішення, що включає факторні навантаження (factor loading, F1) та комунальність (communality, h2).

```
          F1    h2
Item1 0.999 0.9986
Item2 0.998 0.9964
Item3 0.982 0.9647
Item4 0.942 0.8881
Item5 0.934 0.8727
Item6 0.135 0.0182
Item7 0.482 0.2328

SS loadings:  4.972
Proportion Var:  0.71
```

Рис. 5.3.1 Результати факторного рішення

Всі завдання, окрім 6 та 7, в нашому випадку мають дуже міцний зв'язок з латентною здібністю, що вимірюється тестом (Рис. 5.3.1). SS loadings (4.972) – це загальна пояснена дисперсія латентною здібністю. Proportion Var (0.71) – це частка загальної дисперсії, яку пояснює здібність. Тобто, модель має одну суттєву латентну здібність, що пояснює ~71% дисперсії відповідей на завдання.

Однак у IRT нас зазвичай більше цікавлять фактичні параметри завдання, такі як дискримінативність, складність та ймовірність вгадування (нагадуємо, її в нашій моделі немає). Як бачимо, значення параметрів нахилу (a-параметри) коливаються аж від 0.232 до 45.570 (Рис. 5.3.2). Найвищі значення дискримінативності мають завдання 1, 2 та 3, проте і всі інші завдання мають доволі високу розрізнювальну здатність. Параметри

складності показують, що загалом завдання є дуже легкими: на них очікувано дадуть правильну відповідь навіть студенти з доволі низькими рівнями здібностей. Якщо дивитися на ці параметри одночасно, то можна зробити висновок, що більшість завдань є відносно легкими (окрім шостого пункту), але при цьому непогано диференціюють зовсім слабких студентів (окрім шостого пункту): чітко відділяють тих, хто взагалі не має базового рівня здібностей, тобто з подальшою модифікацією їх можна залишити в тесті, щоб одразу розрізняти студентів з найнижчим рівнем знань.

	a	b1
Item1	45.570	-1.695
Item2	28.307	-1.197
Item3	8.896	-1.163
Item4	4.796	-1.582
Item5	4.457	-1.357
Item6	0.232	2.952
Item7	0.938	-1.252

Рис. 5.3.2 Параметри завдань побудованої моделі

Криві характеристик завдань та криві інформації завдань для покрокового тесту на основі STACK наступні (Рис. 5.3.3 та Рис. 5.3.4).

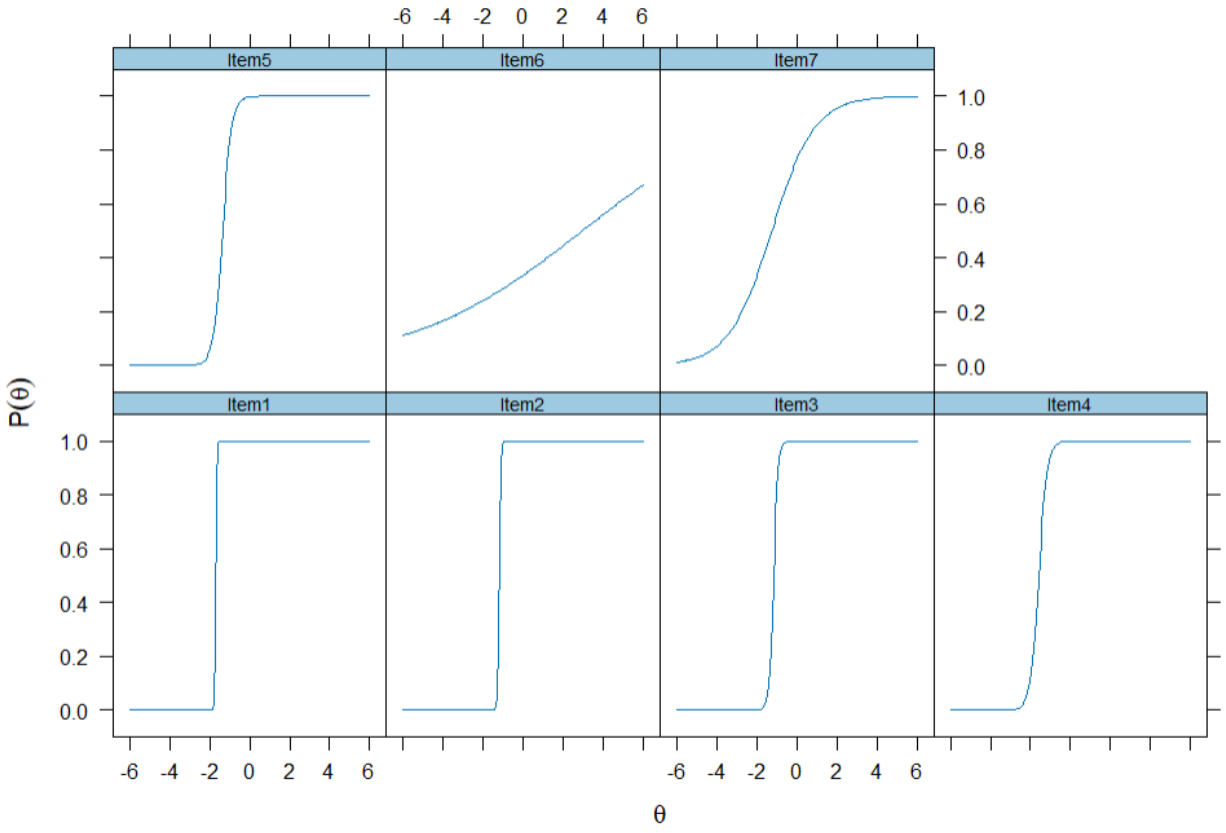


Рис. 5.3.3 Криві характеристик завдань

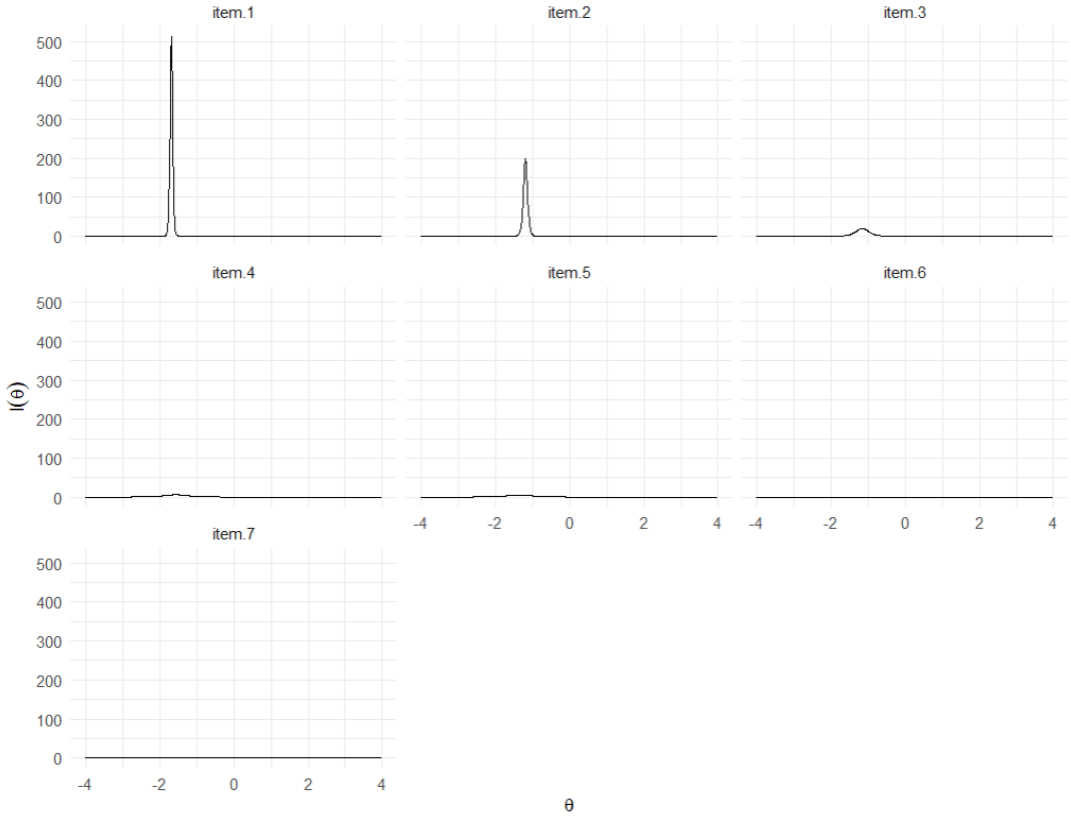


Рис. 5.3.4 Криві інформації завдань

Далі оцінимо, як добре модель відповідає наявним даним. Як бачимо, статистика M2 є значущою (Рис. 5.3.5), що свідчить про суттєві відмінності між побудованою моделлю та зібраними даними, але варто зауважити, що це може відбуватися через відсутність варіацій у вибірці: майже всі студенти отримали максимальний бал.

	M2	df	p	RMSEA	RMSEA_5	RMSEA_95	SRMSR	TLI	CFI
stats	27.24366	14	0.01789593	0.1036809	0.04177029	0.1604598	0.09403688	0.9508946	0.967263

Рис. 5.3.5 Статистики відповідності побудованої моделі

Почнемо з перевірки відповідності завдань (Рис. 5.3.6): як ми бачимо, для всіх завдань, окрім 7, статистика не могла бути обчислена через відсутність варіацій у відповідях – майже всі відповіді були однакові та правильні. Отже, отримані результати вказують на доцільність перегляду тесту. Також бачимо, що завдання 3, 4, 5, 6 та 7 поводяться стабільно у своїй зоні здібностей θ, всі інші завдання є менш продуктивними, але не спотворюють тест. Показник outfit оптимальний лише для 7 та 6 завдання, для інших він занижений, але моделі ці пункти тесту не шкодять (Рис. 5.3.7).

	item	S_X2	df.S_X2	RMSEA.S_X2	p.S_X2
1	Item1	NA	NA	NA	NA
2	Item2	NA	NA	NA	NA
3	Item3	NaN	0	NaN	NaN
4	Item4	NA	NA	NA	NA
5	Item5	NaN	0	NaN	NaN
6	Item6	NaN	0	NaN	NaN
7	Item7	0.784	1	0	0.376

Рис. 5.3.6 Статистики відповідності отриманих параметрів завдання

	item	outfit	z.outfit	infit	z.infit
Item1		0.000	NaN	0.000	NaN
Item2		0.002	-9.009	0.018	-2.223
Item3		0.110	-10.638	1.068	0.298
Item4		0.226	-2.139	1.053	0.264
Item5		0.234	-0.949	1.374	1.344
Item6		0.986	-0.162	0.989	-0.129
Item7		0.828	-1.119	0.941	-0.383

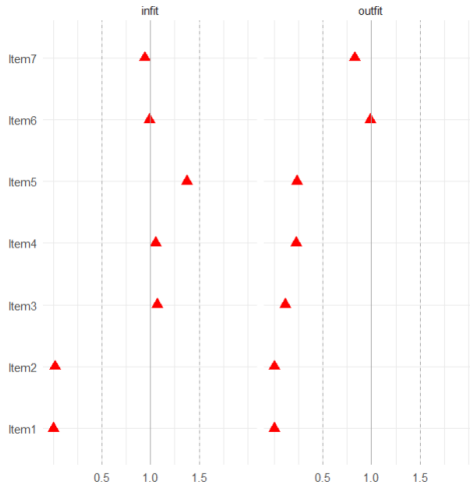


Рис. 5.3.7 Статистики (infit-outfit) відповідності отриманих параметрів

Оцінка відповідності особи (Рис. 5.3.8): як бачимо, для infit та outfit статистик взагалі немає спостережень, які знаходяться за межами інтервалу оптимальних значень, отже більшість респондентів демонструють узгодженість відповідей з побудованою моделлю (можливо, тест є дуже простим для цих студентів).

infit.outside	outfit.outside
1	1

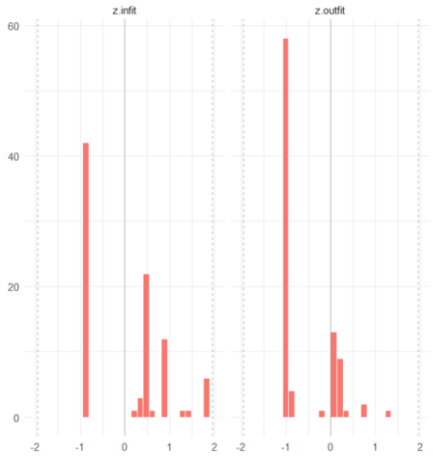
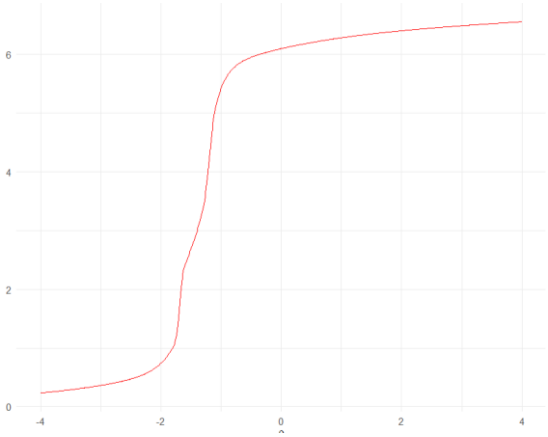


Рис. 5.3.8 Статистики (infit-outfit) відповідності здібностей особи

Останньою характеристикою якості IRT-моделі є графік характеристичної кривої шкали та звичайне обчислення кореляції між латентним балом IRT та спостережуваним (Рис. 5.3.9). Як бачимо, між оцінками спостерігається висока кореляція (0.88)




```
        Pearson's product-moment correlation

data:  score and sumscore
t = 17.851, df = 87, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8315266 0.9240068
sample estimates:
      cor
0.8863016
```

Рис. 5.3.9 Оцінка відповідності побудованої моделі

5.4 Загальні висновки щодо якості тесту на основі STACK

Загалом отримані результати вказують на те, що тест на основі STACK демонструє ознаки задовільної калібровки в межах розглянутої вибірки, хоча більшість завдань виявилися досить легкими для студентів даної вибірки. Це не дивно, адже студенти ФІОТ завжди вирізнялися високим рівнем знань та здатністю вирішувати навіть найскладніші завдання. Попри це, аналіз виявив цікаві закономірності щодо кожного пункту тесту.

Завдання 1 та 4 виявилися досить легкими для студентів, більшість змогли правильно виконати їх. Водночас вони мають низьку дискримінативну здатність, тобто не дуже добре розрізняють студентів за рівнем знань. Це свідчить про те, що ці завдання виконують більше навчальну, ніж диференційну функцію – вони перевіряють засвоєння базових концепцій, але не дозволяють чітко відокремити слабких від сильних учнів.

Завдання 2, 3 та 5 продемонстрували оптимальний баланс між легкістю і здатністю розрізняти рівні підготовки студентів. Вони достатньо зрозумілі, щоб більшість студентів могла отримати правильну відповідь, але при цьому

дозволяють диференціювати середньо і гарно підготовлених студентів, адже в них з’являються варіації у виконанні.

Останні два завдання тесту є найскладнішими. Вони добре розрізняють студентів із різним рівнем підготовки, адже лише кращі виконують їх правильно. Саме ці завдання забезпечують високий рівень дискримінативності і дають тесту розрізнювальний потенціал, дозволяючи виділити тих, хто опанував матеріал глибше, від тих, хто знає його поверхнево.

Взаємна кореляція між завданнями свідчить про логічну узгодженість тесту – всі завдання перевіряють споріднені навички, але при цьому останні завдання більш незалежні і додають цінної інформації для диференціації студентів.

Загалом можна зробити висновок, що тест має високу внутрішню узгодженість та логічну структуру. Хоча для більшості студентів він виявився дещо простим, що свідчить про високий рівень академічної підготовки учасників. Завдання добре описують одну спільну латентну характеристику – математичну компетентність, що свідчить на користь адекватності побудованої моделі в межах даного дослідження.

Розділ 6. Порівняльний аналіз стандартного покрокового тесту та покрокового тесту на основі STACK

Розглянувши властивості кожного тесту окремо, ми тепер порівняємо їхні психометричні характеристики та практичні аспекти.

Стандартний покроковий тест був побудований з 9 варіантами, тоді як тест на основі STACK використовував лише один шаблон. Однак один шаблон STACK може генерувати 704 унікальні задачі, що забезпечує більшу варіативність для тесту, хоча й з складнішою конструкцією. Також стандартні тести можуть вимагати багаторазового коригування можливих варіантів відповідей після тестування для належної роботи з ручним розширенням. На відміну від цього, тест на основі STACK використовує символьну перевірку відповідей, яка значною мірою зменшує потребу в таких діях та робить перевірку розв'язків студентів більш швидкою та гнучкою. Крім того, символьна перевірка відповідей тесту на основі STACK та вбудований попередній перегляд допомагають зменшити кількість помилок друку у студентів.

Як стандартний покроковий тест, так і тест на основі STACK демонструють високу загальну успішність студентів, про що свідчить концентрація балів у верхніх квартилях. Для стандартного тесту 50% балів знаходяться між квартилями $Q_1 = 0.68$ та $Q_3 = 0.95$, тоді як для тесту на основі STACK 50% студентів мають загальний бал вище 0.96. В обох випадках коробкові діаграми показують невелику кількість «викидів» низьких балів, що може свідчити про наявність кількох студентів з нижчим рівнем підготовки, або взагалі з її відсутністю.

Як бачимо, обидва тести демонструють унімодальні розподіли з лівою асиметрією, але тест на основі STACK демонструє більш виражену асиметрію (-3.09 проти -1.45) та набагато вищий ексцес (9.36 проти 1.8), що вказує на гостріший пік та важчі хвости (Рис. 6.1). Це говорить про те, що тест на основі STACK у межах даної вибірки виявився легшим, ніж стандартний тест,

оскільки більше студентів досягають найвищих балів, що можна пояснити можливістю попереднього перегляду та символьною перевіркою. Обидва тести не проходять тести на нормальність (Шапіро-Вілка та Колмогорова-Смірнова). Тест на основі STACK у межах досліджуваної вибірки продемонстрував вищу внутрішню узгодженість, ніж стандартний тест (0.9 проти 0.77 для α Кронбаха та 0.91 проти 0.78 для ω Макдональда).

Показник	Стандартний тест	Тест з STACK
Вибіркове середнє	0.7724	0.8864
Медіана	0.8429	0.9643
Мода	0.9600	0.9600
Стандартне відхилення	0.2273	0.2146
Коефіцієнт асиметрії	- 1.4537	- 3.0906
Коефіцієнт ексцесу	1.7973	9.3358
Тест на нормальність К-С	0.0032	2.8628E-09
Тест на нормальність Ш-В	3.5809E-08	1.5489E-15
Середня кореляція завдання-завдання	0.3292	0.5789
α Кронбаха	0.7699	0.8978
ω Макдональда	0.7814	0.9095

Рис. 6.1

В обох тестах перші чотири завдання були відносно легкими та демонстрували низьку дискримінативність, тоді як останні три (для тесту на основі STACK ці завдання теж мали високий рівень легкості, що виходив за межі оптимального інтервалу, проте в порівнянні з іншою частиною тесту вони були найскладнішими) зробили найбільший внесок у диференціацію продуктивності (Рис. 6.2). Тест на основі STACK загалом показує дещо вищі середні бали для більшості пунктів, але його індекси дискримінативності для всіх пунктів, окрім перших двох, є значно нижчими, ніж у стандартному тесті. Слід також зазначити, що тест на основі STACK має вищі значення кореляції завдання-тест та завдання-решта майже для всіх пунктів тесту. Важливо, що найвищі показники кореляції у тесті на основі STACK спостерігаються для завдань 1-4, які хоч і є легшими ніж останні 3 завдання тесту, проте є критичним для правильного їхнього вирішення. Цього не спостерігається у стандартному тесті, що може свідчити про те, що тест на основі STACK більш адекватно відображає логічну структуру розв'язування в межах даного дослідження.

Комірка	Дискримінативність		Легкість		Кореляція завдання-тест		Кореляція завдання-решта	
	Стандартний тест	Тест з STACK	Стандартний тест	Тест з STACK	Стандартний тест	Тест з STACK	Стандартний тест	Тест з STACK
u	0.0870	0.1600	0.9639	0.9551	0.4416	0.8739	0.3392	0.8342
dv	0.1739	0.3100	0.8855	0.9129	0.5613	0.8404	0.4582	0.7750
du	0.3478	0.3300	0.7651	0.9073	0.4531	0.8184	0.2809	0.7443
v	0.5435	0.2400	0.8253	0.9326	0.7850	0.8169	0.6622	0.7475
uv	0.9348	0.3600	0.6928	0.8989	0.7599	0.8444	0.5861	0.7681
vdu	0.7935	0.4800	0.6235	0.7500	0.7652	0.7259	0.6216	0.6162
int	0.6630	0.4600	0.6506	0.8483	0.7159	0.6707	0.5490	0.5270

Рис. 6.2

Тест на основі STACK демонструє кращі показники внутрішньої узгодженості, структурної відповідності до розв’язування задач та практичної ефективності. Порівняно із стандартною версією у межах проведеного дослідження тест на основі STACK пропонує більшу гнучкість, масштабованість та відповідність логіці вирішення задач порівняно зі стандартним тестом. Він вимагає менше попередньо створених варіантів, ефективніше виконує перевірку символьних відповідей та зменшує ручне втручання під час оцінювання. Однак його розробка є складнішою та вимагає ретельного проектування дерев рішень. З психометричної точки зору обидва тести мають добру надійність у межах розглянутої вибірки, але тест на основі STACK демонструє дещо вищу внутрішню узгодженість та структуру, яка краще відображає основний процес розв’язання проблеми.

Враховуючи всі отримані результати, можна зробити кілька наступних рекомендацій: для тесту на основі STACK доцільно зберегти наявні завдання, особливо ті, що найкраще розрізняють рівні підготовки студентів, проте можна розглянути можливість введення диференційованого зважування завдань залежно від їхніх рівнів складності та дискримінативності, також бажано залучати більш різноманітну вибірку студентів (включаючи тих, хто має нижчий рівень підготовки) для покращення репрезентативності результатів і точності оцінки завдань; для стандартного покрокового тесту слід переглянути легкі завдання та підвищити їх дискримінативність, наприклад, через додаткові перевірки знань або збільшення рівня складності.

У підсумку, обидва підходи до тестування є ефективними для оцінювання рівня знань студентів, але формат на основі STACK демонструє низку переваг у межах даного дослідження, зберігаючи при цьому дещо вищу

надійність. Завдяки цілеспрямованим удосконаленням він має потенціал слугувати надійним та ефективним інструментом оцінювання математичних здібностей студентів за умови подальшого удосконалення. Стандартний тест все ще залишається корисним інструментом, але його використання потребує додаткових зусиль для підтримки високої точності та диференціації студентів.

Висновки

У магістерській дисертації проведено комплексне дослідження якості покрокових комп'ютерних тестових завдань з математичних дисциплін на основі методів класичної теорії тестів та сучасної теорії тестування. Робота спрямована на аналіз структури покрокових тестів, особливостей збору й підготовки емпіричних даних, а також інтерпретацію результатів статистичного аналізу з урахуванням обмежень застосовуваних підходів.

У результаті проведеного дослідження отримано такі основні висновки:

1) Проведений аналіз наукових джерел з класичної та сучасної теорії тестування показав, що методи класичної теорії тестів залишаються доцільними для первинного статистичного аналізу якості тестових завдань, однак мають обмеження у випадку покрокових тестів зі складною структурою, зокрема через припущення незалежності завдань та залежність показників від характеристик вибірки.
2) Встановлено, що покрокові тестові завдання з математичних дисциплін відрізняються структурною неоднорідністю, оскільки окремі кроки можуть мати різну складність і різний внесок у розрізнення студентів за рівнем підготовки. Це зумовлює необхідність аналізу якості не лише тесту загалом, а й кожного окремого кроку.
3) Здійснено ручний збір результатів тестування студентів та виконано підготовку емпіричних даних до статистичного аналізу, що включало очищення, кодування та перевірку коректності даних. Такий підхід забезпечив можливість подальшого коректного застосування методів CTT та IRT.
4) За результатами аналізу в межах класичної теорії тестів оцінено показники складності, дискримінативності та надійності покрокових тестових завдань. Отримані значення дозволили виявити кроки з надмірною або недостатньою складністю, а також обмежену

варіативність результатів, що впливало на інтерпретацію показників якості тестів.

5) Застосування моделей IRT дало змогу здійснити більш детальний аналіз психометричних характеристик окремих кроків тестів та оцінити відповідність моделей емпіричним даним. Показано, що моделі IRT є ефективним інструментом аналізу покрокових тестів, однак їх застосування потребує обережної інтерпретації результатів з огляду на обсяг вибірки та можливі залежності між кроками.
6) Порівняльний аналіз результатів, отриманих методами CTT та IRT, підтвердив доцільність їх поєднання для комплексного оцінювання якості покрокових тестових завдань. Класична теорія тестів забезпечує базовий описовий аналіз, тоді як IRT дозволяє глибше дослідити властивості окремих завдань і тесту загалом.
7) Отримані результати можуть бути використані для вдосконалення підходів до аналізу якості покрокових тестових завдань у системі Moodle з використанням модуля STACK.

Таким чином, поставлену в магістерській дисертації мету досягнуто, а всі завдання дослідження виконано. Проведене дослідження підтверджує доцільність використання статистичних методів класичної та сучасної теорії тестування для аналізу якості покрокових тестів та визначає напрями подальших досліджень, пов’язаних із розвитком моделей аналізу тестів зі складною структурою.

Список літератури

1. Hussein A., Gasmalla H. E. Introduction to the Psychometric Analysis. // In: Gasmalla H. E. E. (ed.) Written Assessment in Medical Education. – Springer, 2023. – P. 112–115.
2. Muñiz J. Test Theories: Classical Theory and Item Response Theory // Papeles del Psicólogo. – 2010. – Vol. 31, No. 1. – P. 57–66.
3. Muñiz J., Bartram D. Improving Personnel Selection through the Use of IRT // International Journal of Selection and Assessment. – 2007. – Vol. 15, No. 1. – P. 58–63.
4. Jimam N. S., Ahmad S., Ismail N. E. Psychometric Classical Theory Test and Item Response Theory Validation of Patients' Knowledge, Attitudes and Practices // Journal of Young Pharmacists. – 2019. – Vol. 11, No. 2. – P. 186–191.
5. Bijlsma H., Rollett W., Spiel C. A Reflection on Student Perceptions of Teaching Quality from Three Psychometric Perspectives: CCT, IRT and GT // In: Rollett W. et al. (eds.) Student Feedback on Teaching in Schools. – Springer, 2021. – P. 42–45.
6. DiCerbo K., Behrens J. Impacts of the Digital Ocean on Education // Technology, Knowledge and Learning. – 2014. – Vol. 19, No. 1. – P. 85–99.
7. DiCerbo K. Psychometric Methods: Theory into Practice // Measurement: Interdisciplinary Research and Perspectives. – 2019. – Vol. 17, No. 1. – P. 60–64.
8. Gannon T. A., Keown K., Rose M. An Examination of Current Psychometric Assessments of Child Molesters' Offense-Supportive Beliefs // International Journal of Offender Therapy and Comparative Criminology. – 2008. – Vol. 53, No. 3. – P. 316–333.
9. Hardin E. E., Leong F. T. Decision-Making Theories and Career Assessment: A Psychometric Evaluation of the Decision Making Inventory // Journal of Career Assessment. – 2004. – Vol. 12, No. 1. – P. 22–41.

10. Meguellati S., Samia A., Ferhat A., Djelloul A., Khalifa Z. A Critical Analysis of the Use of Classical Test Theory (CTT) in Psychological Testing: A Comparison with Item Response Theory (IRT) // Pakistan Journal of Life and Social Sciences. – 2024. – Vol. 22, No. 2. – P. 9442–9449.
11. Kaplan R. M., Saccuzzo D. P. Psychological Testing: Principles, Applications and Issues. – Pacific Grove: Brooks Cole Pub. Company, 1997.
12. Henning G. A Guide to Language Testing: Development, Evaluation, Research. – Cambridge, Mass.: Newbury House Publishers, 1987.
13. Zubairi A. M., Kassim N. L. A. Classical and Rasch Analysis of Dichotomously Scored Reading Comprehension Test Items // Malaysian Journal of ELT Research. – 2006. – Vol. 2. – P. 1–20.
14. Kline D. Chapter 5: Classical Test Theory: Assumptions, Equations, Limitations, and Item Analyses [Електронний ресурс] // Psychological Testing: A Practical Approach to Design and Evaluation. – Режим доступу:https://www.psycholosphere.com/Kline_Chapter_5_Classical_Test_Theory.pdf
15. Hambleton R.K., Swaminathan H. Item response theory: Principles and applications / R.K. Hambleton, H. Swaminathan. — Boston : Kluwer-Nijhoff Publishing, 1985.
16. Embretson S.E., Reise S.P. Item response theory / S.E. Embretson, S.P. Reise. — London : Psychology Press, 2013.
17. Van der Linden W.J., Hambleton R.K. (eds.) Handbook of modern item response theory / W.J. Van der Linden, R.K. Hambleton. — New York : Springer, 1997.
18. DeMars C. Item Response Theory / C. DeMars. — Cary, NC : Oxford University Press, 2010.
19. Ayala R.J. The theory and practice of item response theory / R.J. Ayala. — Reference and Research Book News, 2009. — Vol.24, No.2.

20. Li Y., Baron J. Behavioral Research Data Analysis with R / Y. Li, J. Baron. — New York : Springer, 2012. — Chapter 8.
21. Lord F.M. Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability // Psychometrika. — 1983. — Vol.48. — P.233–245.
22. Lord F.M. Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory // Journal of Educational Measurement. — 1986. — Vol.23, No.2. — P.157–162.
23. Stone C.A. Recovery of Marginal Maximum Likelihood Estimates in the Two-Parameter Logistic Response Model: An Evaluation of MULTILOG // Applied Psychological Measurement. — 1992. — Vol.16, No.1. — P.1–16.
24. Green D.R., Yen W.M., Burket G.R. Experiences in the application of item response theory in test construction // Applied Measurement in Education. — 1989. — Vol.2, No.4. — P.297–312.
25. Da Rocha N.S., Chachamovich E., de Almeida Fleck M.P., Tennant A. An introduction to Rasch analysis for Psychiatric practice and research. — 1879-1379. — Available here.
26. Cook K.F., O'Malley K.J., Roddey T.S. Dynamic assessment of health outcomes: time to let the CAT out of the bag? // Health Services Research. — 2005. — Vol.40, No.5 Pt 2. — P.1694–1711.
27. Edwards M.C. An Introduction to Item Response Theory Using the Need for Cognition Scale // Social and Personality Psychology Compass. — 2009. — Vol.3, No.4. — P.507–529.
28. Rizopoulos D. ltm: An R package for latent variable modeling and item response theory analyses // Journal of Statistical Software. — 2006. — Vol.17, No.5. — P.1–25.
29. Masur P.K. How to run IRT analyses in R / P.K. Masur. — May 13, 2022. — URL:https://philippmasur.de/2022/05/13/how-to-run-irt-analyses-in-r/

30. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов./под общей ред. В.И. Звонникова и М.Б. Челышковой. – М.: Логос.-2012.- 688с.
31. Кухар Л.О. Конструювання тестів. Курс лекцій: навч. посіб. / Л.О.Кухар, В.П.Сергієнко.- Луцьк, 2010.-182с.
32. Лісова Т.В. Моделі та методи сучасної теорії тестів / Т.В.Лісова.- Ніжин: Видавець ПП Лисенко М.М., 2012.-112с
33. Алєксєєва І.В. Покрокові тести з курсу «Методи математичної економіки» / І.В. Алєксєєва, І.В. Орловський, Ю.В. Сорокіна // Вісімнадцята міжнародна наукова конференція ім. акад. Михайла Кравчука.-7-10 жовтня,2017.-Луцьк-Київ.-т.2.- С.168-170
34. Орловський І. В., Тимошенко О. А. Аналіз якості нових методів контролю знань з вищої математики в технічному університеті / І.В. Орловський, О.А. Тимошенко // Шоста міжнародна науково-практична 54 конференція «Математика в сучасному технічному університеті» 28—29 грудня, 2017.- Київ - С. 374-377.
35. Авраменко О.В., Павличенко Г.Ю., Паращук С.Д. Статистичні методи в освітніх вимірюваннях. Частина І. Класична теорія тестування: Навчально-методичний посібник.– Кіровоград: Лисенко В.Ф., 2012.- 120 с.
36. Лісова Т. В. Моделі та методи сучасної теорії тестів : навч.-метод. посіб. / Т. В. Лісова. – Ніжин : НДУ ім. М. Гоголя, 2012. – 111 с.

Додаток 1

Програмний код в R для аналізу якості покрокових тестів

```
# --- 1. Підключення необхідних бібліотек ---
required_pkgs <- c("dplyr", "psych", "mirt", "moments", "ggplot2", "reshape2")
new_pkgs <- required_pkgs[!(required_pkgs %in% installed.packages()[,"Package"])] if(length(new_pkgs)) install.packages(new_pkgs) lapply(required_pkgs, library, character.only = TRUE)
# --- 2. Імпорт матриці з оцінками за кожне завдання ---
data <- read.csv("C:\\Users\\user\\Desktop\\folder\\step-by-step test results.csv", header = TRUE, sep = ";", dec = ".", stringsAsFactors = FALSE)
# --- 3. Розрахунок сумарного балу для кожного студента ---
data$row_total <- rowSums(data, na.rm = TRUE)
items <- data %>% select(starts_with("Item"))
item_names <- colnames(items)
n_items <- ncol(items)
n_students<-nrow(items)
scores <- data$row_total/n_items
# --- 4. Візуалізація (гістограма + діаграма скриня+вуса) ---
par(mfrow = c(1, 2), mar = c(5, 4, 4, 2) + 0.1)
hist(scores, main = "Гістограма", xlab = "Загальна оцінка", ylab = "Частота", breaks = seq(0, 1, by = 1/n_items), col = "royalblue", border = "black", ylim = c(0, 35))
boxplot(scores, main = "Діаграма скриня-вуса", ylab = "Загальна оцінка", col = "royalblue", border = "black", ylim = c(0, 1) )
par(mfrow = c(1, 1))
quantile(scores)
# --- 5. Загальні описові показники ---
mean_score <- mean(scores, na.rm = TRUE)
median_score <- median(scores, na.rm = TRUE)
get_mode <- function(x) {ux <- unique(x) ux[which.max(tabulate(match(x, ux)))]}
```

```
mode_score <- get_mode(round(scores, 2))
sd_score <- sd(scores, na.rm = TRUE)
skewness_score <- moments::skewness(scores, na.rm = TRUE)
kurtosis_score <- psych::kurtosi(scores, type = 1)
# --- 6. Кореляційні показники  ---
item_test_corr <- sapply(item_names, function(it) cor(items[[it]], scores, use = "pairwise.complete.obs"))
item_rest_corr <- sapply(item_names, function(it) {
  total_wo <- (scores*n_items - items[[it]])/n_items
  cor(items[[it]], total_wo, use = "pairwise.complete.obs")})
item_item_corr_matrix <- cor(items, use = "pairwise.complete.obs")
sd_item <- sapply(items, sd, na.rm = TRUE)
corr_melt <- melt(item_item_corr_matrix)
ggplot(corr_melt, aes(Var1, Var2, fill = value)) + geom_tile() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) + labs(title = "Матриця кореляцій між завданнями", x = "", y = "")
aver_item_corr <-mean(item_item_corr_matrix[upper.tri(item_item_corr_matrix)], na.rm = TRUE)
# --- 7. Індекси  ---
item_difficulty <- colMeans(items, na.rm = TRUE)
top_pct <- 0.27
n_top <- ceiling(n_students * top_pct)
ord <- order(scores, decreasing = TRUE)
top_idx <- ord[1:n_top]
bot_idx <- ord[(n_students - n_top + 1):n_students]
disc_index <- sapply(item_names, function(it) {mean(items[[it]][top_idx], na.rm = TRUE) - mean(items[[it]][bot_idx], na.rm = TRUE)})
alpha_res <- psych::alpha(items, check.keys = TRUE)
cronbach_alpha <- alpha_res$total$raw_alpha
omega_res <- psych::omega(items, nfactors = 1, plot = FALSE)
```

```
omega_total <- omega_res$omega.tot
#  --- 8.Перевірка на нормальність ---
shapiro_p <- shapiro.test(scores)$p.value
ks_p <- tryCatch(ks.test(scale(scores), "pnorm")$p.value, error = function(e) NA)
# --- 9. IRT  ---
mod_grm <- tryCatch(mirt(items, 1, itemtype = "graded", verbose = FALSE), error = function(e) NULL)
summary(mod_grm)
irt_params <- if(!is.null(mod_grm)) coef(mod_grm, IRTpars = TRUE, simplify = TRUE) else NULL
M2(mod_grm)
itemfit(mod_grm)
library(ggmirt)
itemfitPlot(mod_grm)
data %>% count(Item1) %>% arrange(desc(n))
personfit(mod_grm)
personfitPlot(mod_grm)
personfit(mod_grm) %>%
  reframe(infit.outside = prop.table(table(z.infit > 1.96 | z.infit < -1.96)),
          outfit.outside = prop.table(table(z.outfit > 1.96 | z.outfit < -1.96)))
if(!is.null(mod_grm)) plot(mod_grm, type = "trace", which.items = 1:n_items)
itemInfoPlot(mod_grm, facet = T)
testInfoPlot(mod_grm, adj_factor = 2)
person_scores <- fscores(mod_grm)
hist(person_scores, main = "Item–Person Map", xlab = "Latent Trait (θ)", col = "lightblue", freq = FALSE)
scaleCharPlot(mod_grm)
sumscore <- rowSums(data)
cor.test(person_scores, sumscore)
```