**Київський національний університет імені Тараса Шевченка**
**Факультет комп'ютерних наук та кібернетики**
**Кафедра системного аналізу та теорії прийняття рішень**

**КВАЛІФІКАЦІЙНА РОБОТА**

на здобуття ступеня бакалавра
за освітньо-професійною програмою «Системний аналіз»
за спеціальністю 124 «Системний аналіз»
на тему:

# МОДЕЛІ І МЕТОДИ АДАПТИВНОГО ТЕСТУВАННЯ ЗНАНЬ

**студента 4 курсу**
**Кушнаренка Івана Сергійовича**

**Науковий керівник:**
**асистент, кандидат технічних наук**
**Махно М. Ф.**

**Робота заслухана на засіданні кафедри системного аналізу та теорії прийняття рішень та рекомендована до захисту в ЕК,**

**протокол №**10 **від** 07.06. ***2022*** **р.**

**Завідувач кафедри системного аналізу та теорії прийняття рішень**
**проф. Наконечний О.Г.**

**Київ – 2022**

# ЗМІСТ

# ВСТУП

Із початком застосування комп'ютерів у навчальному процесі особлива увага приділялася контролю знань. Технічні засоби навчання насамперед використовувалися саме для перевірки знань учнів. І до теперішнього часу, незважаючи на бурхливий розвиток різних форм комп'ютерного навчання, тестувальні програми становлять половину наявних в Інтернеті програм навчального призначення (універсальні та спеціалізовані навчальні системи, електронні енциклопедії, навчальні ігри тощо) і є найбільш опрацьованими.

Проблеми комп'ютерного контролю знань (КЗ) зазвичай розглядаються у двох аспектах: методичному та технічному.

До методичних аспектів належать:

- планування та організація проведення контролю;
- визначення типів питань та відбір завдань для перевірки знань студентів;
- формування набору питань та завдань для опитування;
- визначення критеріїв оцінки виконання кожного завдання та контрольної роботи в цілому.

До технічних аспектів належать:

- автоматичне формування набору контрольних завдань на основі вибраного підходу;
- вибір та використання у системі контролю заданих параметрів оцінювання;
- вибір алгоритмів для оцінки знань учнів.

Тому питання комп'ютерного контролю знань цікавлять багатьох вчених, а саме педагогів і фахівців у сфері інформаційних технологій.

Особливо **актуальною** проблема комп’ютерного контролю знань постає сьогодні, у час глобальної діджиталізації, зокрема навчального процесу. До того ж, в умовах дистанційного навчання, можуть стати в нагоді технології адаптивного тестування знань, які дозволяють оцінювати учнів

більш об’єктивно, виключаючи суб’єктивні фактори оцінювання учнів як, наприклад, персональне ставлення викладача до учнів.

**Мета роботи** полягає у співставленні адаптивних та неадаптивних методів тестування знань і визначенні переваг адаптивних методів.

**Завданням роботи** є: порівняльний аналіз адаптивних та неадаптивних методів тестування знань, визначення особливостей та переваг вищезгаданих методів та їх практичне підтвердження.

**Об’єкт роботи** – процес тестування знань.

У роботі застосовуються такі **методи дослідження** як: порівняльний аналіз, наукові закони та моделювання процесу тестування знань на реальних даних. Порівняльний аналіз та наукові закони використані для аналітичної частини, а моделювання процесу тестування знань для практичної реалізації.

**Можливими сферами застосування** можуть бути: навчальний процес та підтвердження знань.

# 1. Існуючі методи проведення контролю знань

Процес контролю знань складається з трьох етапів: формування питань на основі контрольних завдань; видача їх студенту та отримання його відповіді, можливо, із зворотним зв'язком; виставлення оцінки за контроль. Перші два етапи відносяться до організації процесу комп'ютерного контролю знань.

Методи організації контролю знань можна поділити на три класи:

- неадаптивні методи;
- частково адаптивні методи;
- повністю адаптивні методи.

Далі розглянемо кожен із наведених класів окремо.

## 1.1. Неадаптивні методи контролю

Спільним для всіх неадаптивних методів є те, що варіант контрольної роботи для кожного студента формується до контролю, що, з одного боку, підвищує швидкість контролю (не потрібен пошук завдання у БД та його завантаження), з іншого – дозволяє видавати завдання двома способами: одним або списком. В останньому у разі студент сам може вибрати послідовність виконання завдань.

До неадаптивних методів контролю належать:

*1. Сувора послідовність.*

Набір завдань для контролю заздалегідь підготовляється викладачем або розробником контрольної роботи і міститься в БД системи. Як правило, це однакова послідовність запитань для всіх студентів. Недоліки цього методу очевидні: відсутність різноманітності, що є однією з вимог педагогіки, зниження самостійності виконання завдань та ін. Цей метод вважається найгіршим, незважаючи на широку розповсюдженість. Метод можна трохи покращити, наприклад, підготувавши кілька варіантів

контрольної роботи та/або видаючи завдання студентам у довільної послідовності.

2. *Випадкова вибірка.*

Набір завдань формується безпосередньо перед контролем з урахуванням завдань, тобто із загального набору питань випадковим чином обираються *n* питань для кожного студента. Це значення може бути заздалегідь задано викладачем або випадково обрано студентом. Перевага даного методу полягає в тому, що кожному студенту пропонується індивідуальна послідовність питань. Основний недолік методу – варіант контрольної роботи генерується без урахування складності завдань. Таким чином, набір завдань для одного студента може включати лише найважчі питання, а іншого – лише легкі. Це часто призводить до спотворення результатів контролю. Існують різні модифікації даного методу, що дозволяють враховувати метадані запитань. Наприклад, можуть бути задані тема та загальний час контролю, час відповіді на кожне запитання, кількість спроб дати відповідь тощо або встановлюється кількість питань різного ступеня.

*3. Комбінований метод*

В його основі метод “Випадкова вибірка”, доповнений "Суворою послідовністю". В цьому випадку викладач задає один або кілька питань, які неодмінно повинні бути включені до кожного варіант контрольної роботи. Інші завдання генеруються випадковим чином, як у другому методі.

## 1.2. Частково адаптивні методи

Процес управління адаптивного контролю знань можна зобразити як складну систему. Схематична діаграма цієї системи наведена нижче на рис.1.

*Блок "Алгоритм контролю"* виконує перевірку правильності відповідей студента та виконуваних ним дій, управління процесом контролю знань на

основі обраного методу та визначення результатів контролю, яке зазвичай зводиться до виставленню оцінки студенту.

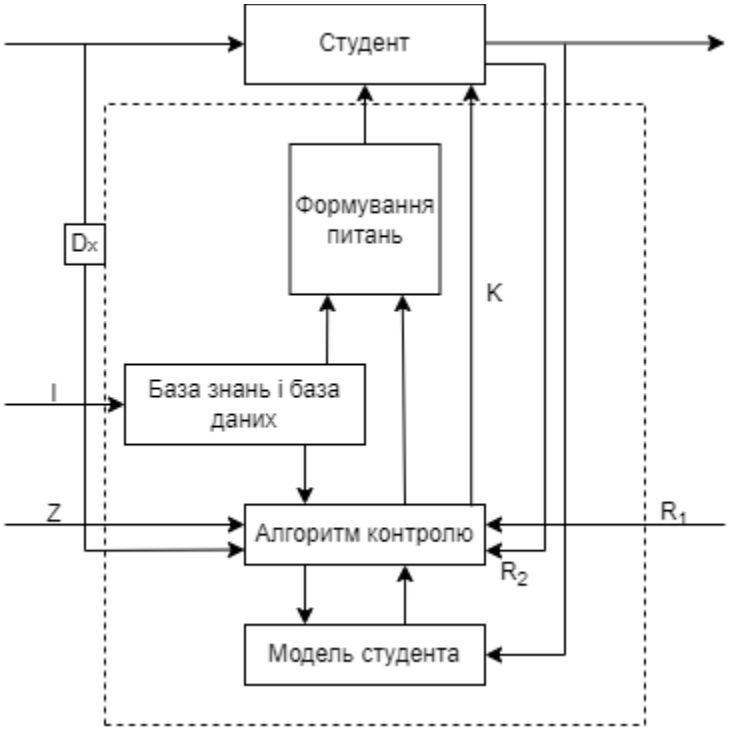


Рис. 1. Модель адаптивного контролю знань

*База знань (БЗ)* містить методи та/або моделі процесу контролю, а також сукупність знань предметної галузі. *База даних (БД)* включає набори питань та завдань, призначених для перевірки знань студента та/або дані для формування завдань. Контрольні завдання можуть також генеруватися автоматично з урахуванням БЗ. База даних та база знань спільно з моделлю студента утворюють репозиторій системи контролю.

*Модель студента* включає різноманітну інформацію про студента: передісторію навчання, результати поточної роботи, особистісні психологічні характеристики, загальний рівень підготовленості.

*Блок «Формування питань»* використовується для формування та видачі студенту чергового завдання.

Контроль знань здійснюється так: студент виконує запропоноване завдання, і результат роботи поміщається в модель студента. Блок "Алгоритм контролю" на основі аналізу відповіді студента, цілей контролю Z та використовуваного методу проведення контролю з огляду на зовнішні ресурси $R_1$ (наприклад, можливості системи контролю) та внутрішні ресурси студента $R_2$ (наприклад, час контролю), а також стан середовища $D_x$ визначає параметри завдання, яке має бути запропоноване студенту. Формувальник питань та задач, отримавши від "Алгоритму контролю" дані про параметри

наступного завдання, вибирає з БД та/або БЗ необхідну інформацію І, формує текст завдання та видає його студенту. У найпростішому випадку робота блоку зводиться до вибору потрібного питання чи завдання із бази даних. За деяких видів контролю може бути передбачений зворотній зв'язок К, що полягає у видачі коментаря на відповідь студента.

Частково адаптивні методи контролю передбачають використання інформації з моделі студента або навчального матеріалу (що він зберігається у базі знань при формуванні набору контрольних питань.

*1. Випадкова вибірка з урахуванням окремих параметрів моделі студента.*

Метод є розвитком неадаптивних методів. Він є аналогічним "Випадковій вибірці" та "Комбінованому методу", тобто. набір завдань також формується безпосередньо перед контролем, але при генерації випадкової послідовності завдань використовуються такі параметри Моделі студента, як загальний рівень підготовленості, здатність до навчання тощо. Таким чином, кожному студенту генерується набір завдань, що відповідає його рівню підготовленості та здібностям, що є головною перевагою даного методу. Інша перевага методу: студент, виконуючи завдання, що відповідають його здібностям, не відчуває зайвого психологічного навантаження під час контролю.

В якості недоліків даного методу можна відзначити, що студенти отримують завдання різної складності (це, безумовно, має бути враховано під час виставлення оцінки), тобто. один виконує тільки прості завдання, а інший –лише складні. Тому, генеруючи питання студенту, відповідні його здібностям, доцільно включити в набір один - два завдання підвищеної складності.

*2. Контроль з урахуванням відповідей студента.*

У цьому методі контроль здійснюється за заздалегідь складеним сценарієм або іншими словами, за розгалуженою контролюючою програмою. Приклад такого сценарію наведено на рис. 2, де вершини графа $B_i$

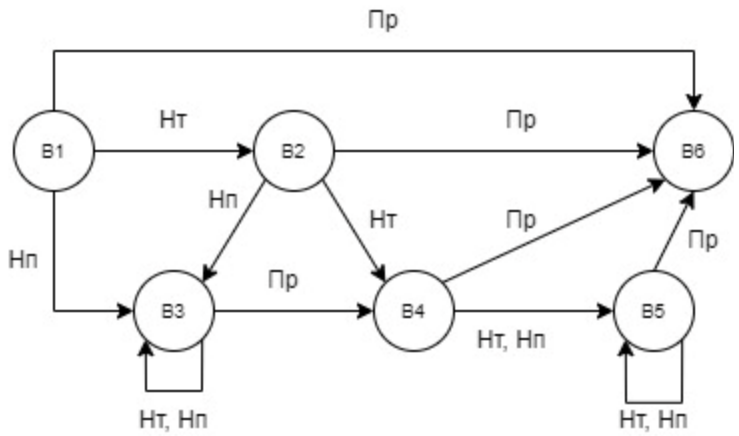


Рис. 2 Сценарій контролю знань

відповідають питанням, запропонованим студенту, а дуги вказують наступний питання, що видається, залежно від правильності відповіді: Пр – правильна відповідь, Нт – неточна, Нп – неправильна відповідь. Попередня підготовка сценарію контролю дає можливість включити програму питання різного ступеня складності та значущості, розташовуючи найбільш значні та важкі завдання в основній галузі програми (на рис. 2 це питання $B_1$ та $B_6$), а простіші – в розгалуженнях. Таким чином, студенти отримують різне число питань, а, отже, і час, що витрачається ними на контроль, є різним, що є перевагою даного методу. Інша перевага методу – простота забезпечення зворотного зв'язку (видачі відповідного коментаря).

Цей метод має один істотний недолік: усім студентам пропонуються одні й ті самі завдання, одного разу включені в контрольну програму. Усунути цей недолік досить просто – достатньо відокремити сценарій видачі питань від набору контрольних завдань. Для цього необхідно підготувати комплект однотипних питань для кожного $B_i$, включеного до сценарію контролю, тобто. $B_i = \{b_{i1}, b_{i2}, \ldots, b_{ik}\}$, а в процесі контролю випадковим чином генерувати студенту питання з комплекту $B_i$.

*3. Контроль на основі моделі навчального матеріалу.*

У цьому методі формування набору завдань для контролю знань відбувається на основі моделі навчального матеріалу, що представляє орієнтований граф: множина вершин графа відповідає об'єктам вивчення, а множина дуг – зв'язкам між ними. Вивчення навчального матеріалу, так само як і організація контролю, здійснюється згідно з оптимальною послідовністю викладу навчального матеріалу, яка зазвичай є ніщо інше, як лінійна

послідовність об'єктів вивчення. Таким чином, спочатку генерується завдання для перевірки знань першого навчального об'єкта, потім другого і так далі, тобто. послідовність видачі завдань аналогічна послідовності вивчення навчального матеріалу. При цьому, якщо планується перевірити і знання, і вміння, то по одному навчальному об'єкту можуть ставитися кілька запитань. Можлива модифікація даного методу, що передбачає генерацію контрольних завдань з урахуванням рівня підготовленості студента.

## 1.3. Адаптивні методи

Адаптивні методи максимально використовують інформацію із моделей студента та навчального матеріалу. Ці методи мають переваги у підвищенні точності вимірювання; оскільки після кожного вибору питання оновлюється інформація про здібності екзаменованого.

*1. Item Response Theory (IRT)*

Теорія відповідей на питання (Item Response Theory, IRT) - це вивчення оцінки тестів та питань, заснованих на припущеннях щодо математичного взаємозв'язку між здібностями екзаменованого (або іншими гіпотетичними характеристиками) та відповідями на запитання. Алгоритм IRT націлений на надання інформації про функціональну залежність між оцінкою рівня студента та ймовірністю того, що студент дасть правильну відповідь на конкретне запитання.

Метод IRT розраховує ймовірність кожного студента правильно відповісти на кожне запропоноване запитання із заданим рівнем складності, враховуючи рівень здібностей студента як параметр функції, який потребує оцінки. Після завершення тесту відповіді студента співвідносяться із обрахованими ймовірностями, що дозволяє оцінити рівень знань студента.

Далі в роботі буде більш детально розглянуто принцип роботи методу IRT.

*2. Комп'ютеризоване адаптивне тестування*

Комп'ютеризоване адаптивне тестування (Computerized Adaptive Testing, CAT) – форма тесту, у якому відповіді студента на попередні запитання впливають на його наступне запропоноване для розв'язку завдання.

Для проведення тесту із використанням цього методу необхідно мати банк запитань, кожне з яких має відповідний рівень складності (так само як у методі IRT). Також треба оцінити початковий рівень знань студента, наприклад, використовуючи дані із попередніх тестувань, визначити його як мінімально допустиме значення для успішного складання тесту, або обрати якесь довільне значення. Власне, останній спосіб не вплине на фінальний результат тестування, оскільки метод CAT є ітераційним, що збігається до певного числа (остаточної оцінки рівня знань студента). Кожне наступне питання, пропонується студенту таким чином, щоб якомога більше уточнити рівень його здібностей.

Тест зупиняється, коли виконується одна із перерахованих умов:

1. Банк питань вичерпано

Це відбувається, як правило, при невеликій кількості підготовлених для тесту питань, коли всі питання вже були запропоновані студенту.

2. Досягнуто максимально допустиму довжина тесту.

Існує заздалегідь встановлена максимальна кількість предметів, які дозволено давати студенту. Зазвичай це та ж кількість предметів, що і в еквівалентному неадаптивному тесті.

3. Показник здібностей оцінюється з достатньою точністю.

Кожна відповідь дає більше статистичної інформації про показники здібностей, підвищуючи точність за рахунок зменшення стандартної помилки.

На m-му кроці алгоритму оновлюється значення параметра $B_m$ (рівень знань студента) за наступною формулою:

$$B_{m+1} = B_m + \frac{R_m - \sum_{i=1}^{m} P_{mi}}{\sum_{i=1}^{m} P_{mi}(1 - P_{mi})}$$, де

$B_k$ – рівень знань студента на кроці $k$;

$R_m$ – кількість правильних відповідей на вже запропоновані $m$ питань;

$P_{mi}$ – ймовірність студента із рівнем знань $B_m$ правильно відповісти на питання $i$.

Тут ймовірність студента правильно відповісти на питання обраховується згідно з моделлю Раша, яка також використовується в методі IRT:

$P_{im} = \frac{1}{1 + e^{-(B_m - D_i)}}$, $D_i$ – складність питання $i$.

Після оновлення рівня здібностей студента $B_{m+1}$ має бути запропоноване наступне питання. Якщо остання відповідь студента була правильною, то видається питання, наступне за рівнем у банку питань. У іншому випадку, коли остання відповідь студента була неправильною, наступне питання за рівнем складності має бути на один рівень складності менше попереднього. Очевидно, питання, що пропонуються не можуть повторюватися.

## 2. Опис алгоритму IRT

Основна ідея методу IRT полягає в оцінці знань студента із урахуванням рівня складності завдань. Для кожного студента *i (i = 1,…I)*, рівень знань якого оцінюється латентним параметром $\theta_i$, визначається функція

$$f(\theta_i) = P(X_j = 1|\theta_i), \text{ де}$$

$X_j$ –відповідь студента і на питання j, незалежні випадкові величини;

$P(X_j=1|\theta)$ – ймовірність, що студент наведе правильну відповідь на питання j;

Далі буде розглянуто 4 IRT моделі для з різними наборами параметрів функції f.

*1. Однопараметрична логістична модель (модель Раша, 1-PL)*

Зв’язок між складністю завдання та рівнем знань студента виражається рівністю:

$$\ln\left(\frac{P_1}{P_0}\right) = \theta - b, \text{ де}$$

$P_1$ – ймовірність правильно відповісти на питання;

$P_0 = 1 – P_1$ – ймовірність неправильно відповісти на питання;

$b$ – рівень складності питання;

$\theta$ – рівень знань студента.

Звідси випливає, що ймовірність студента правильно відповісти на питання дорівнює 0.5, якщо рівень його знань співпадає зі складністю завдання ($\theta = b$).

З наведеної рівності можемо виразити $P_1$, таким чином отримаємо формулу для обрахунку ймовірності правильної відповіді студента на наведене питання.

$$P_1 = P(X_j = 1|\theta_i) = \frac{1}{1 + e^{-(\theta_i - b_j)}}$$

*2. Двопараметрична логістична модель (2-PL)*

У цій моделі додається дискримінаційний параметр $a_j$, міра диференційної здатності питання. На практиці більш високе значення параметра дискримінації означає, що ймовірність правильної відповіді зростає швидше по мірі зростання значення θ. Тобто, при більшому значенні параметра a, необхідний нижчий рівень знань для отримання тієї ж ймовірності правильно відповісти на питання. Якщо правильно відповідь на певне питання дає змогу більш високо оцінити здібності студента, то параметр $a_j$ теж має бути більшим. До того ж, параметр не може бути від'ємним, оскільки це означатиме, що студенти, чий рівень здібностей нижче матимуть більший шанс правильно відповісти на запитання.

Саме дискримінаційний параметр відповідає за «адаптивність» алгоритму IRT. Тобто, якщо два студенти відповіли правильно на однакову кількість, але студент 1 має більше відповідей серед складних питань (параметр *a* для яких є вищим), то студент 1 отримає більш високу оцінку за тест.

Формула обрахунку ймовірностей в цьому випадку має наступний вигляд:

$$P_1 = P\left(X_j = 1 \middle| \theta_i\right) = \frac{1}{1+e^{-a_j\left(\theta_i - b_j\right)}}.$$

Для наочної демонстрації впливу параметра $a_j$ на ймовірність правильної відповіді наведено графік:

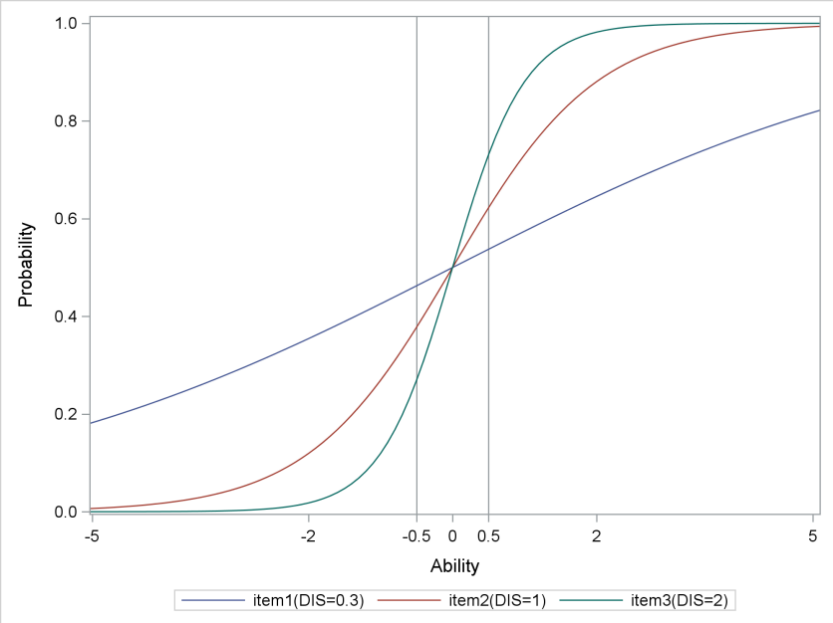


Рис. 3 Залежність ймовірності правильної відповіді на питання з різними значеннями дискримінаційного параметра

*3. Трьохпараметрична логістична модель (3-PL)*

У цій моделі до попередньої додається параметр $c_j$ – коефіцієнт вгадування. Він позначає ймовірність непідготовленого студента вгадати правильну відповідь на наведене питання. Наприклад, у тесті із п'ятьма доступними варіантами відповідей і лише однією правильною параметр $c_j$ буде дорівнювати 0.2. Цей параметр задає мінімальне значення ймовірності правильної відповіді студентом на питання. Для цієї моделі ймовірність рахується за формулою:

$$P_1 = \mathrm{P}\left(X_j = 1 \middle| \theta_i\right) = c_j + \frac{1 - c_j}{1 + e^{-a_j(\theta_i - b_j)}}$$

*4. Чьотирьохпараметрична логістична модель (4-PL)*

Ця модель додатково враховує параметр неуважності студента $d_i$, що задає верхнє значення ймовірності правильної відповіді. Але ця модель рідко використовується на практиці через складність в оцінці чотирьох додаткових параметрів. Формула має наступний вигляд:

$$P_1 = \mathrm{P}\left(X_j = 1 \middle| \theta_i\right) = c_j + \frac{d_i - c_j}{1 + e^{-a_j(\theta_i - b_j)}}$$

На рис. 4 показана характеристична крива чотирьохпараметричної моделі IRT, де b – точка, в якій ймовірність правильної відповіді на питання дорівнює 0,5. a – асимптота, кут нахилу якої відносно осі x відображає диференційну здатність наведеного питання; c – асимптота, що характеризує коефіцієнт вгадування; d – асимптота, що характеризує коефіцієнт неуважності.

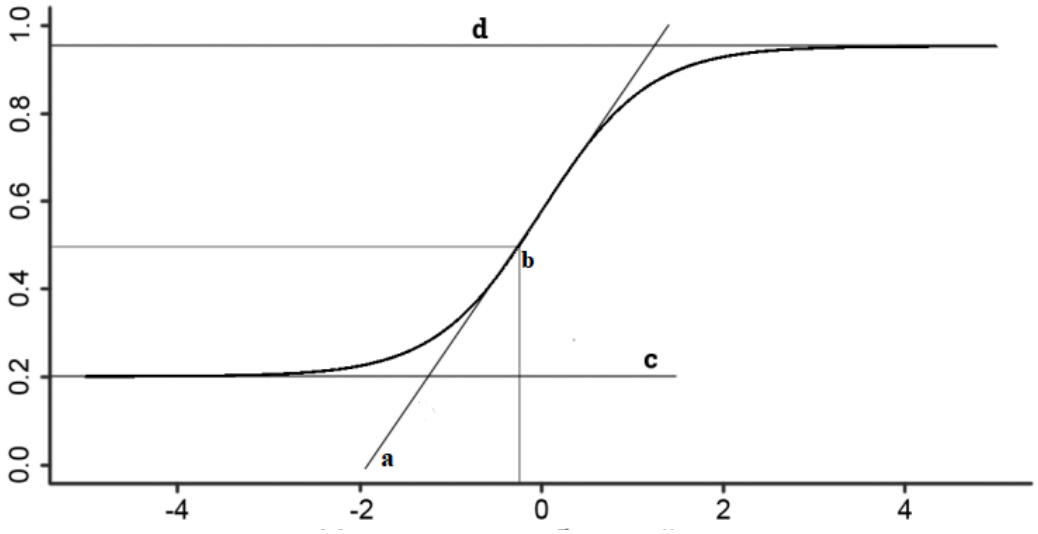


Рис. 4 Крива чотирьохпараметричної моделі IRT

Фінальним етапом методу IRT є оцінка параметру здібностей студента $\theta$. Після отримання відповідей на усі питання тесту, відповіді студента використовуються для отримання кінцевого значення $\theta$, яке буде вважатися результатом роботи (із поправкою на нормування шкали).

Для оцінки параметру $\theta$ зазвичай використовується метод максимальної правдоподібності (ММП). Функція правдоподібності буде мати вигляд:

$$L(\theta_i|\overline{X_i}) = \prod_{k=1}^{N} P_{ik}^{X_k}(1 - P_{ik})^{1-X_k}, \text{ де}$$

$N$ – кількість питань у тесті;

$\overline{X_i} = (X_1, X_2, \dots, X_N)$ – бінарний вектор відповідей студента на питання тесту;

$P_{ik}$ – ймовірність правильної відповіді студентом $i$ на запитання $k$, порахована за однією із чотирьох моделей IRT.

Спрямовуючи функцію $L(.)$ до максимуму, знаходимо оцінку параметра $\theta$.

Але використання ММП має певні недоліки: оцінку знайти неможливо, якщо усі відповіді студента на тест однакові (вектор відповідей $X_i$ має вигляд $(0, 0, ..., 0)$ або $(1, 1, ..., 1)$), а також цей метод краще себе показує на вибірках великого розміру.

## 3. Демонстрація роботи методів адаптивного тестування знань на базі технології IRT

Технології адаптивного тестування знань мають значну кількість переваг над класичними тестами, що було описано вище у роботі. Далі мною буде продемонстровано на практиці принцип роботи 2 моделей технології IRT у порівнянні з тестами із класичною системою оцінювання результатів і буде порівняно результати оцінки знань студентів в обох зазначених методах.

Для демонстрації роботи методів було використано дані із датасету, що надається компанією StataCorp і є доступним за посиланням. Ці дані містять відповіді восьмисот студентів на тест із 9-ма питаннями. Мною було обрано вибірку із 25 студентів. Відповіді студентів представляються бінарним вектором розміру $n$, де $n = 9$ – кількість питань у тесті. Використана вибірка в табличному вигляді розташована у Додатку А.

Отож, для тестування було обрано наступні значення:

кількість питань у тесті $n = 9$;

кількість студентів $I = 25$.

Банк питань був розбитий на 3 класи складності, і кожен з класів оцінюється параметром складності $b_j$ (j = 1..3), що приймає значення 0, 1, 2 відповідно. Також кожному питанню у тесті було поставлено у відповідність значення дискримінаційого параметру $a_j$ (j = 1..n), яке приймає певні значення із діапазону 0..2. Питання із більшим рівнем складності мають більші значення дискримінаційного параметру: $a_1 = 0$, $a_n = 2$, а усі проміжні значення рівномірно розподілені на інтервалі (0, 2).

Для оцінки ймовірності студента відповісти правильно на поставлене питання буде використовуватися модель Раша (1-PL) та двохпараметрична модель (2-PL). Для оцінки параметра знань студента $\theta$ буде використовуватися метод максимальної правдоподібності.

Нагадаю, що функція правдоподібності має наступний вигляд:

$$L(\theta_i|\overline{X_i}) = \prod_{k=1}^{N} P_{ik}^{X_k}(1 - P_{ik})^{1-X_k}.$$

На практиці, якщо є мета знайти максимум функції правдоподібності, зазвичай замість функції правдоподібності розглядається логарифмічна функція правдоподібності. Тоді, розглянемо задачу пошуку максимуму наступної функції:

$$\ln L\left(\theta_i \middle| \overline{X_i}\right) = \ln \prod_{k=1}^{N} P_{ik}^{X_k} (1 - P_{ik})^{1 - X_k} =$$

$$= \sum_{k=1}^{N} \left(X_k ln P_{ik} + (1 - X_k) ln(1 - P_{ik})\right)$$

Для оцінки параметру $\theta$ використовується ітераційний алгоритм, що базується на методі Ньютона-Рафсона.

1. Обирається початкове значення параметра $\theta_i$. Нехай оберемо значення $\theta_i$ = -4, що відповідає найпростішому питанню із запропонованих.
2. Рахуються перша і друга похідна логарифмічної функції правдоподібності в точці $\theta_i$, назвемо ці функції $D_1(\theta_i)$ та $D_2(\theta_i)$.
3. Обраховуємо значення $e = D_1(\theta_i) / D_2(\theta_i)$.
4. Оновлюємо значення $\theta_i := \theta_i - e$.
5. Повторюємо кроки 2-4, допоки абсолютне значення величини $e$ не стане меншим за встановлений поріг (нехай це значення буде дорівнювати 0.001).

$P_{ik}$ – ймовірність студента $i$ правильно відповісти на питання k.

Згідно з моделлю Раша, це значення дорівнює:

$$P_{ik} = \frac{1}{1 + e^{-(\theta_i - b_k)}}$$

Визначимо функції $D_1(\theta_i)$ та $D_2(\theta_i)$:

$$D_1(\theta_i) = \sum_{k=1}^{N} \left(X_k - 1 + \frac{1}{1 + e^{\theta_i - b_k}}\right)$$

$$D_2(\theta_i) = \sum_{k=1}^{N} -\frac{e^{\theta_i - b_k}}{(1 + e^{\theta_i - b_k})^2}$$

А згідно з двохпараметричною моделлю наведені вище функції мають вигляд:

$$P_{ik} = \frac{1}{1 + e^{-a_k(\theta_i - b_k)}}$$

$$D_1(\theta_i) = \sum_{k=1}^{N} a_k \left( X_k - 1 + \frac{1}{1 + e^{a_k(\theta_i - b_k)}} \right)$$

$$D_2(\theta_i) = \sum_{k=1}^{N} -a_k^2 \frac{e^{a_k(\theta_i - b_k)}}{(1 + e^{a_k(\theta_i - b_k)})^2}$$

Мною було розроблено програму на мові Python, задача якої полягала в оцінюванні результатів тестування трьома методами: класичне тестування, для якого результат дорівнює кількості правильних відповідей, IRT із використанням моделі Раша та IRT із використанням двохпараметричної моделі. Після цього, за допомогою ПЗ Microsoft Excel було візуалізовано результати тестів у вигляді гістограм. Код розробленою мною програми розміщено в Додатку Б.

Написана мною програма оцінила результати тестування. Для класичного типу тестування, оцінка видається одразу в процентному вигляді як доля правильних відповідей на тест до загальної кількості питань. Адаптивні методи, розраховують оцінку студента як значення параметра $\theta$. Для порівняння результатів різних видів тестувань та інформативності отриманої студентом оцінки значення $\theta$ треба нормувати до тієї ж шкали, в якій виставлені оцінки при класичному методі тестувань, тобто необхідно конвертувати отримані оцінки здібностей в процентну шкалу 0-100%%.

На перший погляд, для конвертації $\theta$-значень у проценти достатньо визначити відповідність: $\theta^{(0)} \leftrightarrow 0\%$, $\theta^{(1)} \leftrightarrow 100\%$, де $\theta^{(0)}$ – оцінка здібностей студента, який відповів неправильно на усі питання, $\theta^{(1)}$ – оцінка здібностей студента, який на всі питання відповів правильно. Але при проведенні

аналітичних розрахунків виявляється, що ці значення дорівнюють -∞ та +∞ відповідно. Тому, було вирішено провести шкалювання шляхом співвіднесення результатів на двох шкалах.

Якщо студент відповів правильно тільки на найлегше питання у тесті, вважаємо, що він набрав мінімальну можливу кількість балів за тест і в процентній шкалі отримає таку ж оцінку, як і при класичному оцінюванні тесту. У випадку тесту, що складається з 9 питань, при класичному оцінюванні студент би отримав оцінку 1/9 = 11% за правильну відповідь лише на найлегше питання, таку ж оцінку він отримає і при адаптивному підході ($\theta^{(min)} \leftrightarrow 11\%$). Аналогічно, якщо студент *не* відповів правильно лише на найлегше запитання, він отримає таку ж оцінку, як і при класичному тестуванні – 89% ($\theta^{(max)} \leftrightarrow 89\%$). Маючи дві опорні точки на шкалі, ми здатні інтерполювати усі значення $\theta$ на відсотковій шкалі.

Далі наведено графіки із результатами тестувань побудовані у програмі MS Excel на основі даних отриманих за допомогою різних методик оцінок студентів. Ці дані є результатом роботи розробленої мною програми та вказані в Додатку В.

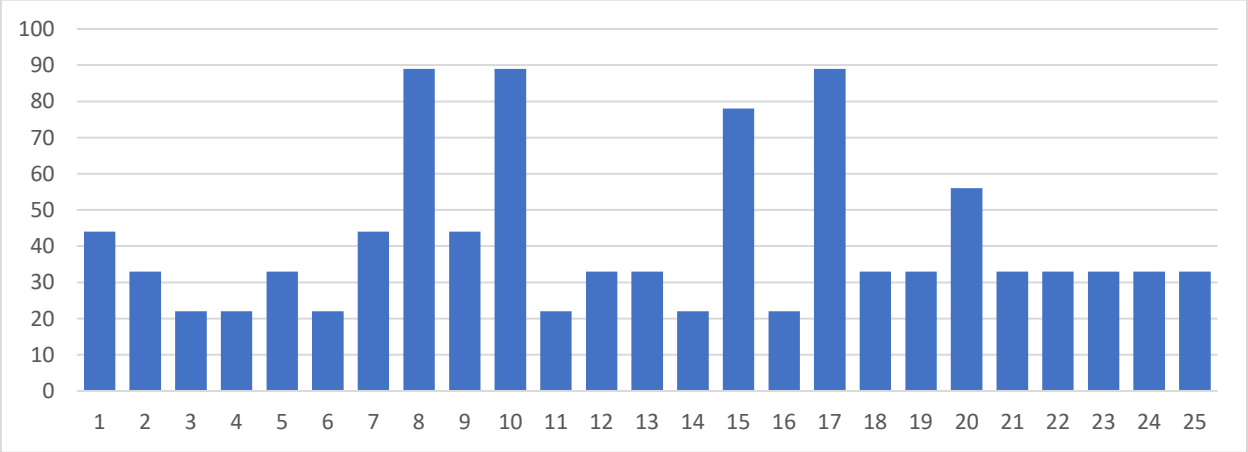


Рис. 5 Результати тестування класичним методом

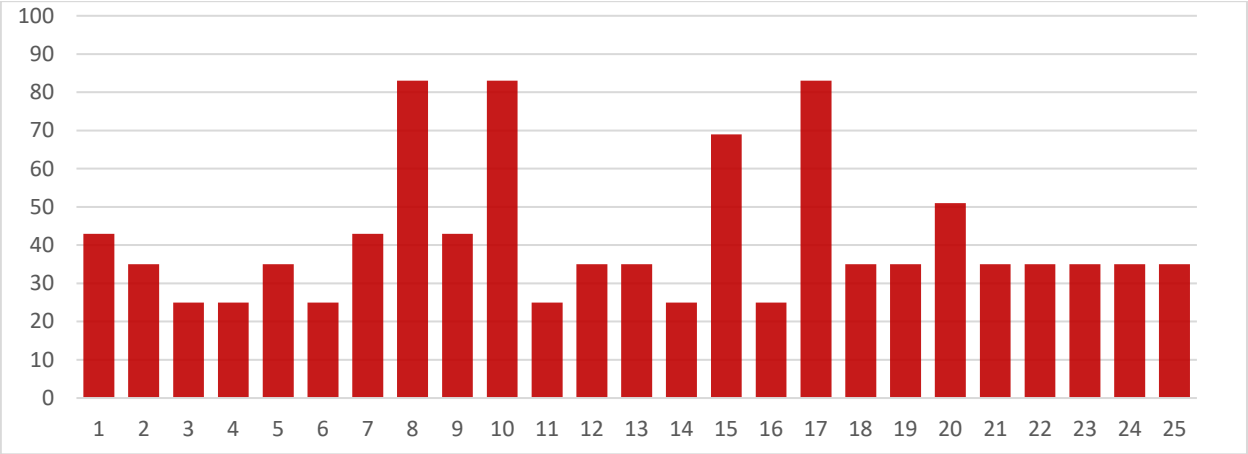


Рис. 6 Результати тестування із моделлю Раша

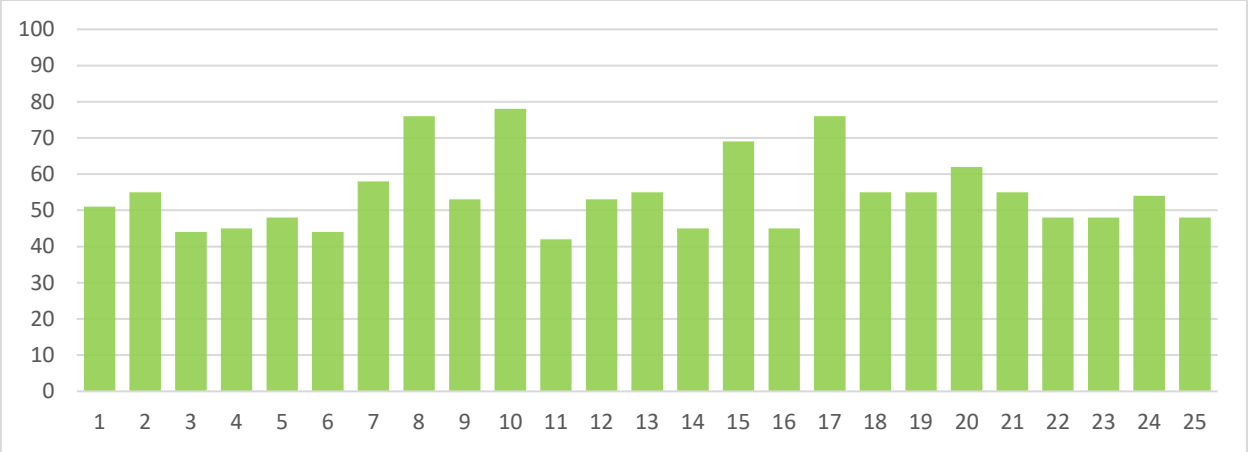


Рис. 7 Результати тестування із моделлю 2-PL

На рис. 5 та рис.6 видно, що метод, що базується на моделі Раша, не вносить покращень у результат тестування: відносно один одного результати кожного зі студентів такі ж самі, як і при застосуванні класичного тестування. Це пов'язано із тим, що модель Раша не враховує дискримінаційний параметр, який і відповідає за «адаптивність» навчання.

Натомість, двохпараметрична модель дуже помітно змінює ситуацію із результатами тестувань. У групі студентів збільшилася кількість різних за абсолютним значенням оцінок, студенти які при класичному підході мали однакові бали, тепер маються різні, результати оцінювання є більш точними, оскільки враховують складність запропонованих питань.

Отже, стає очевидним те, які переваги надає адаптивне тестування у порівнянні із класичним:

1. Більш точні оцінки за результатами тестування

2. Більша різноматність результатів оцінок
3. Об’єктивність отриманих оцінок

Головний недолік, що проявив себе при використанні адаптивного підходу, це менший діапазон оцінок, що в цілому пов’язаний із вибором методу шкалювання результатів адаптивного тестування. Проте, неможливість побудувати звичну нам шкалу результатів не є проблемою у тестах, результат яких визначається критерієм “склав – не склав”, де просто обирається порогове значення параметру $\theta$, яке необхідне для успішного складання іспиту.

# ВИСНОВКИ

У роботі були розглянуті різні методи тестування знань: неадаптивні, частково-адаптивні та адаптивні, вказані недоліки неадаптивних методів, показано, як ці недоліки обходяться в різних адаптивних методах, та описані переваги адаптивних методів перед іншими. Серед цих переваг:

- Більша точність вимірювання
- Більш об’єктивне оцінювання результатів тесту
- Більша різноманітність результатів тестування

Детально було розглянуто метод IRT для адаптивного тестування знань. Моделі IRT враховують параметри питань, такі як складність та дискримінаційність, і видають оцінку після повного складання тесту студентом. В IRT правильна відповідь на більш складне питання забезпечує більш вискоий результат після проходженя тесту.

Також, на наборі тестових даних було проведено демонстрацію роботи методів IRT та проведено порівняння результатів в моделях IRT та класичному тестуванні. В результаті, було доведено, що IRT модель 2-PL має усі переваги адаптивних методів тестування. Проте, IRT модель 1-PL (модель Раша) дає такі ж результати як і неадаптивні методи. Тобто, в рамках IRT модель Раша не є прикладом адаптивного тестування, хоча алгоритм CAT може використовувати модель Раша саме в адаптивному підході.

Отже, адаптивне тестування – дієва альтернатива застарілому підходу до тестування знань. В умовах глобальної комп’ютеризації світу імплементація адаптивних методів стає все більш доступною і вже впроваджується в різних тестових системах, наприклад іспит на знання іноземних мов TOEFL. А різноманітність підходів дозволяє вибрати один із безлічі методів, який найбільше підходить для конкретної поставленої задачі. У адаптивних методів тестування великі перспективи для розвитку і поширення, і їх впровадження безумовно стане значним покращенням у всіх сферах, де використовується тестування знань.

## Список літератури


1. Аванесов В. Item Response Theory: Основные понятия и положения.
2. Мазорчук М. С. Методы и модели анализа качества тестовых заданий и моделирование компьютерного адаптивного тестирования в системах дистанционного обучения / Мазорчук М. С., Добряк В. С., Емельянов П. С. - Национальный аэрокосмический университет им. Н. Е. Жуковского «Харьковский авиационный институт», 2016.
3. Ado Abdu Bichi Item Response Theory: An Introduction to Latent Trait Models to Test and Item Development / Ado Abdu Bichi, Rohaya Talib - Universiti Teknologi Malaysia, Malaysia, 2018.
4. Cees A.W. Glas Item response theory in educational assessment and evaluation - ADMEE-Canada - Université Laval, 2008.
5. Eric Loken Estimation of a four-parameter item response theory model / Eric Loken, Kelly L. Rulison - The British Psychological Society, 2010
6. John Michael Linacre Computer-Adaptive Testing: A Methodology Whose Time Has Come - Seoul, South Korea: Komesa Press, 2000.
7. Jorge N. Tendeiro IRT (GMMSGE01): Parametric IRT (dichotomous data) – 2017.
8. Julie Wood Logistic IRT Models – 2017.
9. Jumailiyah Mahmud Item response theory: A basic concept - Institute of Teaching and Educational Sciences of Mataram, Indonesia, 2017
10. Mansoor Al-A'ali IRT - Item Response Theory Assessment for an Adaptive Teaching Assessment System – Ahila University, 2006.
11. Nonna Shapovalova Adaptive testing model as the method of quality knowledge control individualizing / Nonna Shapovalova, Olena Rybalchenko, Iryna Dotsenko, Svitlana Bilashenko, Andrii Striuk, Levan Saitgareev - Kryvyi Rih National University.
12. Ruslan Jabrayilov Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment / Ruslan Jabrayilov, Wilco H. M.

Emons, Klaas Sijtsma – National Center fot Biotechnolodgy Information, 2016.

13. Sophiana Chua Abdullah Student Modelling By Adaptive Testing - A Knowledge-Based Approach - The University Of Kent At Canterbury, 2003.
14. Si-Mui Sim Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a para-clinical multidisciplinary paper / Si-Mui Sim, Raja Isaiah Rasiah - Ann Acad Med Singap, 2006.
15. Xinming An Item Response Theory: What It Is and How You Can Use the IRT Procedure to Apply It / Xinming An, Yiu-Fai Yung - AS Institute Inc., 2014.
16. Zhongmin Cui Comparison of Algorithms that Allow Item Review in Computerized Adaptive Testing / Zhongmin Cui, Chunyan Liu, Yong He, Hanwei Chen – 2018.

# ДОДАТОК А

## Таблиця тестових даних

| Student id | q1 | q2 | q3 | q4 | q5 | q6 | q7 | q8 | q9 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 4 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 5 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 6 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| 7 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 8 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 9 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 10 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 11 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| 12 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |
| 13 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 14 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 |
| 15 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 16 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 17 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 18 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 19 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 |
| 20 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |
| 21 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 22 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 23 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| 24 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |

# ДОДАТОК Б
## Код програми на мові Python для оцінки результатів тестування

```
import csv

import dataset
from math import exp, fabs


items = 9

discr_lower = 0.5
discr_upper = 2

threshold = 0.001


def item_difficulty_level(item_n: int) -> int:
    ''' item_n is from [0, items) '''
    if not (0 <= item_n < items):
        raise ValueError("item number is out of range")

    if 0 <= item_n < 3:
        return 0;
    elif 3 <= item_n < 6:
        return 1
    elif 6 <= item_n < 9:
        return 2


def item_discrimination(item_n: int) -> float:
    if not (0 <= item_n < items):
        raise ValueError("item number is out of range")

    return discr_lower + (discr_upper - discr_lower) / (items -
1) * item_n



def d1_rasch(t, answers):
    res = 0
    for item_n, response in enumerate(answers):
        exponent = t - item_difficulty_level(item_n)
        res += response - 1 + 1/(1 + exp(exponent))
    return res


def d2_rasch(t, answers):
    res = 0
    for item_n, response in enumerate(answers):
        exponent = t - item_difficulty_level(item_n)
        res -= exp(exponent) / (1 + exp(exponent)**2)
    return res
```

```
def d1_2pl(t, answers):
    res = 0
    for item_n, response in enumerate(answers):
        discrimination = item_discrimination(item_n)
        exponent = discrimination * (t -
item_difficulty_level(item_n))
        res += discrimination * (response - 1 + 1/(1 +
exp(exponent)))
    return res


def d2_2pl(t, answers):
    res = 0
    for item_n, _ in enumerate(answers):
        discrimination = item_discrimination(item_n)
        exponent = discrimination * (t -
item_difficulty_level(item_n))
        res -= discrimination**2 * (exp(exponent) / (1 +
exp(exponent)**2))
    return res


def ability_level_ctt(student_answers):
    return sum(student_answers) / len(student_answers) * 100


def ability_level_rasch(students_answers: list[int]):
    theta = 0.5
    while True:
        d1_theta = d1_rasch(theta, students_answers)
        d2_theta = d2_rasch(theta, students_answers)

        eps = d1_theta / d2_theta
        if fabs(eps) < threshold:
            break
        theta = theta - eps
    return theta


def ability_level_2pl(students_answers: list[int]):
    theta = 0.5
    while True:
        d1_theta = d1_2pl(theta, students_answers)
        d2_theta = d2_2pl(theta, students_answers)

        eps = d1_theta / d2_theta
        if fabs(eps) < threshold:
            break
        theta = theta - eps
    return theta
```

```
students_responses: list[list[int]] = dataset.students_responses

students_result_ctt = []
students_results_irt_rasch = []
students_results_irt_2pl = []

for answers in students_responses:
    students_result_ctt.append(ability_level_ctt(answers))

students_results_irt_rasch.append(ability_level_rasch(answers))
    students_results_irt_2pl.append(ability_level_2pl(answers))


def scale_irt_rasch_results(theta_results: list[float]):
    min_possible_result_answers = [1, *([0] * (items - 1))]
    max_possible_result_answers = [0, *([1] * (items - 1))]
    min_percents =
ability_level_ctt(min_possible_result_answers)
    max_percents =
ability_level_ctt(max_possible_result_answers)

    min_theta = ability_level_rasch(min_possible_result_answers)
    max_theta = ability_level_2pl(max_possible_result_answers)

    res = []
    for theta in theta_results:
        percent_value = (max_percents - min_percents) /
(max_theta - min_theta) * (theta - min_theta) + min_percents
        res.append(round(percent_value))
    return res


def scale_irt_2pl_results(theta_results: list[float]):
    min_possible_result_answers = [1, *([0] * (items - 1))]
    max_possible_result_answers = [0, *([1] * (items - 1))]
    min_percents =
ability_level_ctt(min_possible_result_answers)
    max_percents =
ability_level_ctt(max_possible_result_answers)

    min_theta = ability_level_2pl(min_possible_result_answers)
    max_theta = ability_level_2pl(max_possible_result_answers)

    res = []
    for theta in theta_results:
        percent_value = (max_percents - min_percents) /
(max_theta - min_theta) * (theta - min_theta) + min_percents
        res.append(round(percent_value))
    return res
```

```
scaled_ctt_results = list(map(round, students_result_ctt))
scaled_irt_rasch_result =
scale_irt_rasch_results(students_results_irt_rasch)
scaled_irt_2pl_result =
scale_irt_2pl_results(students_results_irt_2pl)
```

# ДОДАТОК В

## Таблиця із результатами оцінювання студентів за проведеним тестуванням

| Класичний тест | Модель Раша, θ | Модель Раша, % | 2-PL, θ | 2-PL, % |
|---|---|---|---|---|
| 44 | 0,738348274 | 43 | 0,943497379 | 51 |
| 33 | 0,197987797 | 35 | 1,231107134 | 55 |
| 22 | -0,432189666 | 25 | 0,415819502 | 44 |
| 22 | -0,432189666 | 25 | 0,510387114 | 45 |
| 33 | 0,197987797 | 35 | 0,741166334 | 48 |
| 22 | -0,432189666 | 25 | 0,415819502 | 44 |
| 44 | 0,738348274 | 43 | 1,44843903 | 58 |
| 89 | 3,3254071 | 83 | 2,806829117 | 76 |
| 44 | 0,738348274 | 43 | 1,081556629 | 53 |
| 89 | 3,3254071 | 83 | 2,904101334 | 78 |
| 22 | -0,432189666 | 25 | 0,313236458 | 42 |
| 33 | 0,197987797 | 35 | 1,103618335 | 53 |
| 33 | 0,197987797 | 35 | 1,231107134 | 55 |
| 22 | -0,432189666 | 25 | 0,510387114 | 45 |
| 78 | 2,432253999 | 69 | 2,261610383 | 69 |
| 22 | -0,432189666 | 25 | 0,510387114 | 45 |
| 89 | 3,3254071 | 83 | 2,806829117 | 76 |
| 33 | 0,197987797 | 35 | 1,231107134 | 55 |
| 33 | 0,197987797 | 35 | 1,231107134 | 55 |
| 56 | 1,258979943 | 51 | 1,755538061 | 62 |
| 33 | 0,197987797 | 35 | 1,231107134 | 55 |
| 33 | 0,197987797 | 35 | 0,741166334 | 48 |
| 33 | 0,197987797 | 35 | 0,741166334 | 48 |
| 33 | 0,197987797 | 35 | 1,168359906 | 54 |
| 33 | 0,197987797 | 35 | 0,741166334 | 48 |