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

Фізико-математичний факультет

Кафедра математичного аналізу та теорії ймовірностей

«На правах рукопису»
УДК 517.237

До захисту допущено:
Завідувач кафедри
_____ Олег КЛЕСОВ,
«___» ___травня____ 2025 р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-науковою програмою «Страхова та фінансова математика»

зі спеціальності 111 «Математика»

на тему: «Вибір моделі MIRT для аналізу тестів з вищої математики»

Виконав:
студент ІІ курсу, групи ОМ-31мн,
Шимчик Сергій Дмитрович _____

Керівник:
Кандидат фізико-математичних наук, доцент,
Диховичний Олександр Олександрович _____

Рецензент:
Заступник декана факультету радіофізики,
електроніки та комп'ютерних систем
Київського національного
університету ім. Т. Шевченка
канд.фіз.-мат.наук, доцент,
Єфіменко Світлана Володимирівна _____

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

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-наукова програма «Страхова та фінансова математика»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ Олег КЛЕСОВ

« 04 » лютого 2025 р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Шимчику Сергію Дмитровичу

1. Тема дисертації «Вибір моделі MIRT для аналізу тестів з вищої математики», науковий керівник дисертації Диховичний Олександр Олександрович, доцент кафедри математичного аналізу та теорії ймовірностей, затверджені наказом по університету від «21»березня 2025 р. № 1276-с

2. Термін подання студентом дисертації 15 травня 2025 року.

3. Об’єкт дослідження є моделі MIRT для статистичного аналізу тестів з вищої математики.

4. Предмет дослідження є вибір моделі MIRT

5. Перелік завдань, які потрібно розробити:

1) Ознайомитися з літературою та дослідити основні означення та поняття.
2) Опанувати основні моделі та методи MIRT.
3) Дослідити методи дослідницького факторного аналізу (EFA) для вибору розмірності моделі: CD, EKC, HULL, MAP, NEVALSGT1, RAWPAR, SESCREE, SMT.
4) Дослідити та обрати алгоритми оцінювання латентних параметрів моделей.
5) Дослідити та обрати методи перевірки адекватності моделей MIRT.
6) Дослідити програмну реалізацію відібраних алгоритмів та методів у середовищі R.
7) На підставі відібраних алгоритмів і програм провести статистичний аналіз результатів контрольної роботи бакалаврів РТФ.
8) Відібрати найбільш адекватні моделі MIRT.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу: 20 слайдів.

7. Орієнтовний перелік публікацій:

1) Шимчик С. Д., Диховичний О. О., Круглова Н. В. Вибір моделі MIRT для аналізу тестів з вищої математики // Тези XIII Всеукраїнська наукова конференція молодих математиків, 9-10 травня 2025. Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

9. Дата видачі завдання 04 лютого 2025 року.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1.	Ознайомитися з літературою та дослідити основні означення та поняття.	04.02.23-26.02.23	Виконано
2.	Опанувати основні моделі та методи MIRT.	27.02.25-12.03.25	Виконано
3.	Дослідити методи дослідницького факторного аналізу (EFA) для вибору розмірності моделі: CD, EKC, HULL, MAP, NEVALSGT1, RAWPAR, SESCREE, SMT.	13.03.25-25.03.25	Виконано
4.	Дослідити та обрати алгоритми оцінювання латентних параметрів моделей.	26.03.25-14.04.25	Виконано
5.	Дослідити та обрати методи перевірки адекватності моделей MIRT.	15.03.25-23.03.25	Виконано
6.	Дослідити програмну реалізацію відібраних алгоритмів та методів у середовищі R.	24.04.25-29.04.25	Виконано
7.	На підставі відібраних алгоритмів і програм провести статистичний аналіз результатів контрольної роботи бакалаврів РТФ.	30.04.25-04.05.25	Виконано
8.	Відібрати найбільш адекватні моделі MIRT	05.05.25-12.05.25	Виконано

Студент Сергій ШИМЧИК

Науковий керівник Олександр ДИХОВИЧНИЙ

Реферат

Магістерська дисертація: 41 сторінки, 22 першоджерела, 20 слайдів презентації. Робота складається зі вступу, п’яти розділів, висновків та списку використаної літератури.

В дисертаційній роботі досліджується вибір моделі MIRT для аналізу тестів з вищої математики. Основною метою дисертаційного дослідження є вибір адекватної моделі MIRT для аналізу тестів з вищої математики. Об’єктом дослідження є методи EFA, моделі MIRT, оцінка параметрів та перевірка адекватності моделей для аналізу тестів з вищої математики. Предметом дослідження є вибір моделі MIRT.

Перший розділ містить перелік обраних моделей для статистичного аналізу педагогічних тестів.

Другий розділ містить математичні методи EFA попереднього визначення розмірності моделей MIRT.

Третій розділ містить методи оцінювання латентних параметрів моделей MIRT, які використовуються в роботі.

Четвертий розділ містить методи перевірки адекватності моделі.

П’ятий розділ містить статистичний аналіз результатів контрольної роботи з вищої математики бакалаврів РТФ та вибір адекватної моделі для тестів з вищої математики.

Ключові слова: тести з вищої математики, MIRT, EFA, CFA, алгоритми CD, EKC, HULL, MAP, NEVALSGT1, RAWPAR, SESCREE, SMT, EM, критерії AIC, BIC, M2, RMSEA, CFI ,TLI.

Abstract

Master's thesis: 41 pages, 22 primary sources, 20 presentation slides. The work consists of an introduction, five chapters, conclusions and a list of references.

The thesis investigates the choice of the MIRT model for analyzing tests in higher mathematics. The main purpose of the dissertation is to select an adequate MIRT model for analyzing tests in higher mathematics. The object of the study is EFA methods, MIRT models, parameter estimation and model adequacy testing for the analysis of higher mathematics tests. The subject of the study is the selection of the MIRT model.

The first section contains a list of selected models for statistical analysis of pedagogical tests.

The second section contains mathematical methods of EFA for preliminary determination of the dimensionality of MIRT models.

The third section contains methods for estimating the latent parameters of MIRT models used in this paper.

The fourth section contains methods for checking the adequacy of the model.

The fifth section contains a statistical analysis of the results of the control work in higher mathematics in the Faculty of Engineering bachelors and the choice of an adequate model for tests in higher mathematics.

Keywords: tests in higher mathematics, MIRT, EFA, CFA, CD, EKC, HULL, MAP, NEVALSGT1, RAWPAR, SESCREE, SMT, EM, algorithms, AIC, BIC, M2, RMSEA, CFI, TLI criteria.

Зміст

ВСТУП 7
I. Основні математичні моделі MIRT 9
1. Модель GRM 10
2. Модель GPCM 12
3. Модель NRM 13
II. Попереднє визначення розмірності моделі MIRT 15
III. Оцінка параметрів моделі 21
IV. Перевірка адекватності моделей MIRT 22
V. Аналіз контрольної роботи і вибір адекватної моделі MIRT 26
ВИСНОВКИ 39
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ: 40

ВСТУП

Дистанційна освіта в Україні набрала великої популярності в умовах пандемії COVID-19, а згодом стала важливим елементом навчального процесу через агресію РФ проти України. Блокування можливості проводити традиційні очні заняття та забезпечення безпеки студентів під час воєнного стану зробили дистанційне навчання не лише необхідним, а й постійно актуальним. Переваги дистанційної освіти, такі як доступність, гнучкість та можливість для кожного студента навчатися у зручному темпі, сприяли її широкому впровадженню в українській освіті. Однак, одна з головних проблем, з якими стикаються викладачі та студенти, — це контроль якості знань і ефективність тестування студентів у цих умовах.

Для оцінки знань студентів через онлайн-тести важливим є забезпечення високої якості тестових завдань та їх статистичний аналіз. Основним завданням такого аналізу є не тільки визначення рівня знань студентів, а й виявлення слабких місць у тестах, а також адаптація їх під рівень підготовки студентів. У цьому контексті статистичний аналіз тестів стає надзвичайно важливим інструментом для коригування та поліпшення тестових завдань, що сприяє більш точному визначенню рівня знань.

Статистичний аналіз тестів традиційно базується на двох основних підходах: Clasical Test Theory (CTT) [1] та Item Response Theory (IRT) [2]. Перший підхід зазвичай застосовує загальні показники, такі як середній бал тесту, дисперсія відповідей, що дозволяє отримати загальне уявлення про результати тестування. В той же час, IRT дозволяє створити моделі, які точно оцінюють здатність студента відповідати на конкретні завдання залежно від їхніх індивідуальних здібностей. Ідея IRT полягає в тому, що ймовірність правильної відповіді на завдання залежить не тільки від його складності, а й від здібностей студента.

В свою чергу, більш розвиненою версією IRT є Multidimensional Item Response Theory (MIRT) [3], яка дозволяє одночасно враховувати кілька латентних характеристик, таких, наприклад, як математичні навички, аналітичне мислення, здатність до вирішення задач різної складності. Моделі MIRT дають змогу більш тонко визначати, які саме фактори впливають на успішність студента при виконанні завдань тесту.

Загальна схема для аналізу тестів за допомогою моделей MIRT виглядає наступним чином згідно [4]:

1. Проведення дослідницького факторного аналізу (EFA) [5]: на цьому етапі визначається, скільки латентних змінних (факторів) необхідно

для адекватного опису даних. Для цього застосовуються методи, такі як Comparison Data (CD), Empirical Kaiser Criterion (EKC), Hull Method(HULL), Velicer's MAP Test (Minimum Average Partial) (MAP), Number of eigenvalues greater than 1(NEVALSGT1), Raw Parallel Analysis(RAWPAR), Scree Test(SESCREE), Sequential Mean Test(SMT).

2. Оцінка параметрів моделі (CFA) [6]: після визначення розмірності моделі проводиться підтверджувальний факторний аналіз, щоб перевірити, чи відповідає вибрана модель даним. Тут оцінюються латентні параметри моделі, наприклад, дискримінації (яка здатність кожного завдання дискримінувати студентів за здібностями) та порогові значення тощо.
3. Вибір оптимальної моделі на основі критеріїв якості: для оцінки адекватності моделі використовуються такі статистики, як AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), M2 (M Statistics), RMSEA (Root Mean Square Error of Approximation), CFI (Comparative Fit Index) та TLI (Tucker-Lewis Index) [7] [8] [9] [10] [11].

І. Основні математичні моделі MIRT

MIRT (Multidimensional Item Response Theory) [3]— це теорія, яка розширює традиційну IRT (Item Response Theory), додаючи можливість оцінювати кілька латентних змінних одночасно, замість того, щоб обмежуватись лише одним виміром, як це відбувається в класичній IRT. В основі IRT лежить принцип, що ймовірність правильної відповіді на тестове завдання залежить від однієї латентної змінної, наприклад, загального рівня знань або здібностей. Проте цей підхід не завжди достатньо точний, коли йдеться про складні тести, що оцінюють різноманітні аспекти когнітивних здібностей, такі як аналітичне мислення, критичне сприйняття або здатність до вирішення задач із різними рівнями складності.

MIRT розв'язує цю проблему, дозволяючи аналізувати кілька аспектів чи рис, що впливають на результат тесту. Тобто, замість того, щоб обмежуватися одним загальним виміром, MIRT дозволяє враховувати різні характеристики, які можуть впливати на відповідь респондента. Наприклад, у тесті з вищої математики, де оцінюються не лише математичні навички, але й здатність до просторового мислення чи до логічного вирішення задач, MIRT може одночасно моделювати ці аспекти, даючи більш точну та багатогранну оцінку знань і здібностей студента. Це особливо важливо, коли необхідно визначити, як конкретні характеристики або комбінації навичок впливають на успішність тестування.

Моделі MIRT можуть бути застосовані в контексті різноманітних типів завдань, включаючи завдання з кількома варіантами відповідей або відкриті питання, що дає змогу аналізувати не лише правильність відповідей, але й різні рівні складності або часткові оцінки. Це значно розширює можливості оцінювання і дозволяє глибше вивчати різні аспекти знань або вмінь, що тестуються. Враховуючи цей багатовимірний підхід, MIRT стає потужним інструментом для створення тестів, які можуть вимірювати різні характеристики одночасно, а не просто загальний рівень знань.

Серед моделей MIRT, які використовуються для аналізу тестових завдань, було відібрано моделі: GRM (Graded Response Model), GPCM (Generalized Partial Credit Model) та NRM (Nominal Response Model) [7]. Кожна з цих моделей має свої особливості, які дозволяють враховувати різні аспекти завдань і відповідей. GRM використовується для завдань з кількома категоріями відповідей, де кожен рівень відповіді має свій поріг складності. Наприклад, для завдань, що передбачають кілька варіантів відповіді, ця модель дозволяє оцінити ймовірність того, що респондент вибере певну категорію в залежності від його здібностей. GPCM, у свою чергу, є більш універсальним і дозволяє працювати з завданнями, де можливі часткові бали, що особливо корисно для завдань з відкритими питаннями. Нарешті, NRM використовується для завдань, де варіанти відповідей

не мають чіткої впорядкованості, наприклад, для питань з множинним вибором, де кожен варіант відповіді є рівнозначним.

1. Модель GRM

Модель Graded Response Model (GRM) [7] є потужним інструментом для аналізу тестів, що використовують шкали з кількома впорядкованими категоріями відповідей. Вона найбільше підходить для тестів, у яких відповіді організовані в певному порядку, але не мають конкретних числових значень між категоріями. Прикладами таких завдань можуть бути шкали Лайкерта або інші категоріальні шкали, де респондент обирає відповідь, що найкраще відображає його думку або оцінку: від "погано" до "відмінно", або від "не згоден" до "повністю згоден". Модель GRM дає змогу оцінити ймовірність того, що респондент вибере ту чи іншу категорію залежно від його латентної змінної — рівня знань чи здібностей.

В основі моделі лежить припущення, що ймовірність правильності відповіді залежить не лише від здібностей респондента, а й від складності конкретного завдання a_i (часто називається "параметром дискримінації"). Модель оцінює ймовірність того, що респондент j, чий рівень здібностей позначається через параметр θ_j, вибере певну категорію k або вище для кожного завдання i. При цьому для кожного завдання моделюється не лише загальна ймовірність правильної відповіді, а й ймовірність того, що відповідь потрапить в одну з кількох категорій, розташованих по порядку складності. При цьому в одномірному випадку ($s = 1$) модель використовує наступну формулу для кожного завдання:

$$P\left(X_{ij} \geq k \middle| a_i, d_{ik}, \theta\right) = \frac{1}{1 + \exp\left(-(a_i\theta - d_{ik})\right)},$$

де:

k – рівень категорії;

i – номер завдання;

j – номер іспитника;

a_i – дискримінаційна здатність;

θ_j – рівень підготовленості;

d_{ik} - пороговий рівень категорії k.

У разі, коли латентна змінна θ має векторну природу, тобто здібностей більше ніж одна, наприклад дві, її позначають як $\overrightarrow{\theta_j} = \left(\theta_{j1}, \theta_{j2}\right)$У такій ситуації кожна координата вектора відповідає окремій рисі або здатності респондента.

Наприклад, θ_{j1} може відображати рівень аналітичного мислення, а θ_{j2} — здатність до обчислень.

Дискримінаційний параметр завдання також набуває вигляду вектора $\overrightarrow{a_i} = (a_{i1}, a_{i2})$, де a_{i1} показує, наскільки завдання i чутливе до першої здібності, а a_{i2} — до другої.

Тоді у двовимірному випадку ($s = 2$) ймовірність того, що респондент із такими здібностями обере категорію k або вище для завдання i, обчислюється за формулою:

$$P\left(X_{ij} \geq k \middle| \overrightarrow{a_i}, d_{ik}, \overrightarrow{\theta_j}\right) = \frac{1}{1 + exp\left(-\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2} - d_{ik}\right)\right)},$$

де:

k — рівень категорії;

i – номер завдання;

j – номер іспитника;

$\overrightarrow{\theta_j} = \left(\theta_{j1}, \theta_{j2}\right)$ — рівень підготовленості;

$\overrightarrow{a_i} = (a_{i1}, a_{i2})$ — дискримінаційна здатність;

d_{ik} — пороговий рівень категорії k.

Головною особливістю GRM є використання порогових значень для кожної категорії відповідей. Для кожного завдання визначається ряд параметрів: параметр дискримінації a_i, що характеризує, наскільки добре завдання розрізняє респондентів із різними рівнями здібностей, та пороги b_k, які визначають межу для кожної категорії відповіді. Формула для моделювання ймовірності того, що респондент вибере категорію k у двовимірному випадку ($s = 2$), виглядає так:

$$P\left(X_{ij} = k \middle| \overrightarrow{a_i}, d_{ik}, \overrightarrow{\theta_j}\right) = P\left(X_{ij} \geq k \middle| \overrightarrow{a_i}, d_{ik}, \overrightarrow{\theta_j}\right) - P\left(X_{ij} \geq k - 1 \middle| \overrightarrow{a_i}, d_{ik}, \overrightarrow{\theta_j}\right),$$

$$P\left(X_{ij} = k \middle| \overrightarrow{a_i}, d_{ik}, \overrightarrow{\theta_j}\right) = \frac{1}{1 + \exp\left(-\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2} - d_{ik}\right)\right)} - \frac{1}{1 + \exp\left(-\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2} - d_{i(k-1)}\right)\right)},$$

де:

k — рівень категорії;

i – номер завдання;

j – номер іспитника;

$\overrightarrow{\theta_j} = \left(\theta_{j1}, \theta_{j2}\right)$ — рівень підготовленості;

$\overrightarrow{a_i} = (a_{i1}, a_{i2})$ — дискримінаційна здатність;

d_{ik}, $d_{i(k-1)}$ — порогові рівні категорій k та $k-1$ відповідно.

Тут $a_{i1}\theta_{j1} + a_{i2}\theta_{j2}$ є сумою внесків обох здібностей до результату виконання завдання. Якщо, наприклад, завдання більше вимагає розвинених обчислювальних навичок (θ_{j2}), ніж аналітичного мислення(θ_{j1}), тоді параметр a_{i2} буде більшим за a_{i1}, і внесок θ_{j2} буде визначальним.

Такий підхід дозволяє дуже гнучко і точно описувати багатогранність реальних когнітивних процесів. Модель враховує, що успішність у завданні може залежати не від якоїсь однієї здатності, а від комбінації декількох рис одночасно. Це дає змогу не тільки оцінити загальний рівень знань студента, але й детально проаналізувати сильні та слабкі сторони в окремих напрямах. Саме ця властивість робить GRM у багатовимірному контексті надзвичайно важливим інструментом для якісного тестування.

2. Модель GPCM

Наступною моделлю, яку розглянемо у межах багатовимірного аналізу тестових завдань, є GPCM — Generalized Partial Credit Model [7]. Як і GRM, ця модель працює з політомічними відповідями, тобто із завданнями, для яких передбачено декілька рівнів оцінювання. Проте, на відміну від GRM, де основний акцент робиться на кумулятивній природі переходу між категоріями (тобто вибір "категорії або вище"), у GPCM оцінювання будується безпосередньо на ймовірності вибору кожної окремої категорії.

У GPCM ймовірність вибору певної категорії k для завдання i формується за допомогою логіт-моделі, яка описує сукупний внесок усіх переходів до поточного рівня. Формально ця ймовірність записується так:

$$P\left(X_{ij} = k \middle| \overrightarrow{a_i}, d_{ik}, \overrightarrow{\theta_j}\right) = \frac{exp\left((k-1)\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2}\right) + d_{ik}\right)}{\sum_{l=0}^{k} exp\left((l-1)\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2}\right) + d_{il}\right)}$$

де:

k — рівень категорії;

i – номер завдання;

j – номер іспитника;

$\overrightarrow{\theta_j} = \left(\theta_{j1}, \theta_{j2}\right)$ — рівень підготовленості;

$\overrightarrow{a_i} = (a_{i1}, a_{i2})$ — дискримінаційна здатність;

d_{ik} — пороговий рівень категорії k

У GPCM параметри d_{ik} також можна інтерпретувати як локальні параметри складності, але на відміну від GRM, кожен перехід тут описується незалежно — немає чіткого кумулятивного підходу "категорія або вище". Модель більше нагадує поступове накопичення "балів" чи "кредитів", необхідних для досягнення певної категорії.

Отже, GPCM можна розглядати як природне розширення ідеї дихотомічної логіт-моделі для ситуацій, де відповідь має більше двох можливих рівнів, причому кожен перехід оцінюється окремо та більш гнучко.

3. Модель NRM

Ще однією важливою політомічною моделлю у рамках багатовимірної теорії тестових завдань є NRM — Nominal Response Model [7]. Вона істотно відрізняється від GRM та GPCM за своєю природою, оскільки орієнтується не на упорядковані категорії відповідей, а на номінальні, тобто такі, між якими немає природного порядку. В NRM відповіді на завдання розглядаються як альтернативи, кожна з яких має свою специфічну функцію ймовірності вибору.

Якщо у GRM та GPCM передбачалося, що категорії пов'язані між собою через певну логічну послідовність (категорії "кращі" або "гірші"), то в NRM жодної такої впорядкованості немає. Це особливо корисно для завдань, де варіанти відповіді є якісно різними, наприклад, тест з декількома варіантами правильних або частково правильних відповідей.

Формула для ймовірності вибору категорії k в завданні i для двовимірного випадку ($s = 2$), виглядає так:

$$P\left(X_{ij} = k \middle| \overrightarrow{a_{ik}}, d_{ik}, \overrightarrow{\theta_j}\right) = \frac{exp\left(a_{ik1}\theta_{j1} + a_{ik2}\theta_{j2} + d_{ik}\right)}{\sum_{l=0}^{k} exp\left(a_{il1}\theta_{j1} + a_{il2}\theta_{j2} + d_{il}\right)},$$

де:

k — число рівнів відповідей;

i – номер завдання;

j – номер іспитника;

$\overrightarrow{\theta_j} = \left(\theta_{j1}, \theta_{j2}\right)$ — рівень підготовленості;

$\overrightarrow{a_{ik}} = (a_{ik1}, a_{ik2})$ — дискримінаційна здатність;

d_{ik} — пороговий рівень категорії k.

У цій моделі кожна категорія має свої власні параметри дискримінації і порогові рівні, і вони оцінюються окремо, незалежно від інших. Тобто тут не йдеться про якусь спільну шкалу складності завдання — кожен варіант "бореться" за вибір сам по собі. Багатовимірна природа здібностей дозволяє ще

гнучкіше моделювати ситуації, коли певні відповіді більше пов'язані з однією рисою, а інші — з іншою.

NRM дуже добре підходить, коли необхідно аналізувати багатоваріантні завдання з несхожими між собою відповідями, і коли важливо розуміти не тільки загальну здатність студента, а й його нахил до вибору певного типу відповідей.

Таким чином, GRM, GPCM і NRM дають змогу моделювати різні типи завдань і відповідають різним логікам взаємодії респондента з тестовим матеріалом, а багатовимірність здібностей у всіх цих моделях дозволяє робити аналіз значно тоншим та інформативнішим.

II. Попереднє визначення розмірності моделі MIRT

Визначення розмірності моделей MIRT (Multidimensional Item Response Theory) — це процес визначення, скільки латентних (прихованих) вимірів (факторів) потрібно для адекватного опису взаємозв'язку між відповідями респондентів на завдання тесту. Іншими словами, скільки різних умінь, здібностей або рис вимірює тест.

Первинний аналіз кількості факторів відбувається на основі EFA (Exploratory Factor Analysis) аналізу. Нами були обрані наступні алгоритми.

1. Comparison Data (CD)

Метод Comparison Data [12] бере свої корені з ідеї порівняння власних значень (eigenvalues), отриманих від справжніх даних, з власними значеннями, отриманими від випадкових або згенерованих даних. Його основна мета — визначити, які власні значення з реальних даних є статистично значущими, і, таким чином, допомогти в оцінці кількості факторів.

Ідея цього методу полягає в тому, що для кожної змінної в даних ми можемо провести факторний аналіз і оцінити, скільки факторів "істинно" пояснюють взаємозв'язок між змінними. Власні значення, які виявляються в реальних даних, порівнюються з тим, що ми отримали б, генеруючи випадкові набори даних за допомогою статистичних методів.

Алгоритм виконання методу Comparison Data (CD):

1. Будується кореляційна матриця.
2. Обчислюються власні значення (eigenvalues) для кореляційної матриці.
3. Генерується n (зазвичай близько 1000-5000) випадкових наборів даних з такою ж кількістю змінних та спостережень, як у реальній вибірці.
4. Обчислюються власні значення для кожного випадкового набору даних.
5. Тепер порівнюються власні значення реальних даних з власними значеннями, отриманими від випадкових даних. Для кожного власного значення з реальних даних визначається, чи є воно значущим у порівнянні з випадковими даними.
6. Власні значення, що виявляються значно більшими за ті, що генеруються випадковими даними, можуть бути інтерпретовані як значущі фактори. Це дає кількість факторів, які потрібно залишити для подальшого аналізу.

2. Empirical Kaiser Criterion (EKC)

Метод Kaiser Criterion [13] був запропонований у 1960-х роках як спрощений підхід для визначення кількості факторів у факторному аналізі. Оригінальна версія методу припускала, що фактори з власними значеннями (eigenvalues), більшими за 1, є значущими, оскільки кожен фактор повинен пояснювати більше варіативності, ніж одна змінна.

Empirical Kaiser Criterion (EKC) [14] — це вдосконалена версія методу, яка враховує особливості реальних даних. Вона містить додаткові корекції для забезпечення більш точних результатів, оскільки Kaiser Criterion іноді призводить до надмірного включення факторів. Тобто EKC коригує класичний підхід, використовуючи реальні дані для кращого прогнозування розмірності.

Алгоритм виконання методу Empirical Kaiser Criterion (EKC):

1. Будується кореляційна матриця.
2. Обчислюються власні значення для кореляційної матриці.
3. Застосування стандартного Kaiser Criterion: Тобто, якщо $\lambda_i > 1$, то фактор i є значущим і його варто залишити.
4. Після стандартного Kaiser Criterion застосовують додаткові корекції, щоб врахувати особливості конкретного набору даних. Такі корекції можуть включати в себе порівняння з результатами інших критеріїв, наприклад Scree Test.
5. Якщо корекція вказує на більше число факторів, то це може вказувати на необхідність включення додаткових факторів. Якщо ж корекція вказує на менше число факторів то слід подумати над виключенням деяких факторів.

3. Hull Method (HULL)

Hull Method [15] [16] був запропонований Lorenzo-Seva, Albert Ferrando та іншими дослідниками у 2006 році. Цей метод виник як відповідь на потребу мати більш обґрунтований статистично і менш суб'єктивний спосіб вибору кількості факторів, особливо в контексті моделювання латентних змінних (наприклад, у багатовимірній теорії відповіді на завдання — MIRT).

Основна ідея: знайти модель з найкращим співвідношенням між точністю моделі та її складністю. Тобто не просто модель, яка добре "підходить" до даних, а така, що найефективніша: пояснює якомога більше, використовуючи якомога менше факторів.

Алгоритм виконання методу Hull Method (HULL):

1. Оцінюється кореляційна матриця.
2. Генеруються випадкові дані такої ж розмірності як спостережувані дані. Випадкові дані можуть бути згенеровані з рівномірного або нормального розподілу.
3. Знаходяться власні значення для випадкових даних. Власні значення випадкової матриці позначаються як λ^*.
4. Повторюється пункт 4 багато разів, наприклад, 100 разів, щоб отримати середнє значення власних значень випадкової матриці. Середнє значення власних значень цієї матриці позначається як $E^*\lambda$.
5. Далі порівнюється власне значення λ з відповідним середнім значенням випадкових власних значень $E^*\lambda$. Якщо λ більше за $E^*\lambda$, то фактор

вважається значущим, і на його місце вибирається новий фактор. Кількість факторів збільшується, поки λ не стане меншим за $E^*\lambda$.

6. Остаточна кількість значущих факторів визначається як максимальна кількість факторів, при якій λ більше за відповідне середнє значення випадкових власних значень $E^*\lambda$.

4. Velicer's Minimum Average Partial (MAP) Test

Метод запропоновано Веліцером (Wayne F. Velicer) у 1976 році як відповідь на обмеження методу «eigenvalues > 1». MAP мав на меті дати аналітичну, автоматизовану альтернативу, яка не покладається на суб'єктивність [17].

MAP базується на припущенні, що при правильному числі факторів залишкова (residual) частина кореляційної матриці містить лише шум. Метод поступово "очищає" кореляційну матрицю від факторної структури та шукає той момент, коли середні залишкові часткові кореляції найменші. Якщо ми видаляємо надто мало факторів, у залишковій матриці все ще зберігається значуща факторна структура, що проявляється у високих часткових кореляціях між змінними. Натомість, коли ми вилучаємо занадто багато факторів, починаємо видаляти не лише шум, а й справжню інформацію, і залишкова структура стає переважно випадковою. Отже, оптимальною вважається та кількість факторів, після вилучення якої середні залишкові часткові кореляції є мінімальними.

Алгоритм Velicer's MAP Test:

1. На першому етапі оцінюється кореляційна матриця R.
2. Далі виконується серія факторизацій, кожна з яких поступово видаляє від 0 до $p-1$ факторів. Для кожної фіксованої кількості факторів k здійснюється факторний аналіз (зазвичай PCA або EFA), результатом якого є реконструйована кореляційна матриця R_k. Після цього обчислюється залишкова матриця E_k як різниця між початковою та реконструйованою матрицями:

$$E_k = R - \hat{R}_k$$

На основі залишкової матриці E_k обчислюється матриця часткових кореляцій — тобто кореляцій між змінними, з урахуванням впливу всіх інших змінних. Для кожної пари змінних (i, j), обчислюється квадрат залишкової часткової кореляції $r^2_{ij,k}$.

3. Для кожного значення k (кількість вилучених факторів) обчислюється MAP-індекс як середнє значення квадратів усіх часткових кореляцій у залишковій матриці:

$$MAP_k = \frac{2}{p(p-1)} \sum_{i<j} r^2_{ij,k}$$

Далі порівнюються значення MAP_k для всіх k, і оптимальним вважається таке k, для якого середній квадрат залишкових часткових кореляцій є найменшим. Це означає, що на цьому кроці з моделі вилучено стільки факторів, скільки потрібно для того, щоб залишкова структура була максимально випадковою — тобто без суттєвих латентних зв'язків.

5. Number of eigenvalues greater than 1(NEVALSGT1)

Цей метод запропоновано Г. Кайзером (Henry F. Kaiser) у 1960 році як простий і швидкий спосіб визначення кількості факторів у факторному аналізі. Він став дуже популярним завдяки своїй інтерпретативній простоті й автоматизованості [13].

Метод ґрунтується на тому, що eigenvalue (власне значення) показує, скільки загальної дисперсії у змінних пояснює кожен фактор або компонента. Якщо eigenvalue більше за 1, то цей фактор пояснює більше дисперсії, ніж одна змінна в середньому (оскільки стандартна змінна має дисперсію 1). Це означає, що компонент/фактор є "інформативним" і пояснює більше дисперсії, ніж окрема змінна — отже, його варто залишити.

Алгоритм NEVALSGT1:

1. Будується кореляційна матриця.
2. Обчислюються власні значення для кореляційної матриці.
3. Далі обираються тільки ті фактори, для яких власні значення є більші за 1. Їх кількість і буде вважатися оптимальною розмірністю моделі.

6. Raw Parallel Analysis(RAWPAR)

Метод паралельного аналізу (Parallel Analysis, PA) запропонували Horn (1965) та згодом удосконалювали інші дослідники. RAWPAR — це класичний варіант PA, який працює з сирими (нецентрованими) eigenvalues. Метод став золотим стандартом для визначення кількості факторів завдяки своїй здатності відокремлювати справжню факторну структуру від випадкових флуктуацій у даних [18].

Raw Parallel Analysis полягає в тому, щоб зіставити власні значення (eigenvalues) з реальних даних із тими, які отримуються із випадкових (симульованих) даних, що не мають латентної факторної структури.

Якщо власне значення з реальних даних значно більше, ніж середнє з випадкових даних — це ознака значущого латентного фактора. Якщо ж воно не перевищує той рівень, то, ймовірно, ми маємо справу лише з шумом.

Алгоритм RAWPAR:

1. На початку обчислюється кореляційна матриця R_{real} . Потім обчислюються власні значення $\lambda_1^{real}, \lambda_2^{real}, \dots, \lambda_p^{real}$.

2. Далі симулюється велика кількість (зазвичай 1000) випадкових матриць даних з такою самою кількістю спостережень і змінних, як у вашому наборі. Ці матриці не мають латентної структури (змінні несуміжні, випадкові), але мають однакові розміри.

   Для кожної симульованої матриці обчислюється кореляційна матриця та її eigenvalues. У результаті отримується набір симульованих eigenvalues для кожної компоненти $\lambda_{1,1}^{sim}, \lambda_{1,2}^{sim}, \dots, \lambda_{1,1000}^{sim}$ (перший індекс — номер компоненти, другий — ітерація)
3. Для кожного порядкового eigenvalue обчислюється середнє значення з симуляцій. Потім реальні eigenvalues порівнюються з відповідними середніми симульованими. Компонента або фактор вважається значущим, якщо $\lambda_k^{real} > mean(\lambda_k^{sim})$. Оптимальна кількість факторів — це кількість таких eigenvalues, які перевищують відповідні середні із симуляцій.

7. Scree Test(SESCREE)

Метод запропонував Raymond B. Cattell у 1966 році. Назва "scree" походить від геологічного терміна: *"scree slope"* — кам'яна осип, яка виникає біля підніжжя гори. У цьому методі eigenvalues розглядаються як «скелясті виступи» (інформативні фактори), а далі йде «осип» — дрібні, незначущі компоненти, що відповідають шуму [19].

Scree Test — це візуальний метод, який ґрунтується на побудові графіка eigenvalues у порядку спадання. Після крутої ділянки, де eigenvalues значно зменшуються (інформативні фактори), крива вирівнюється — це зона «осипу». Оптимальна кількість факторів — точка «зламу» або «ліктя» на цьому графіку.

Алгоритм Scree Test:

1. Будуємо кореляційну матрицю R і обчислюємо власні значення $\lambda_1, \lambda_2, \dots, \lambda_p$. Власні значення впорядковують у спадному порядку, щоб побачити, яка частка загальної дисперсії припадає на кожен фактор.
2. Далі будуємо графік eigenvalues, де по осі x — номер компоненти (1, 2, 3…), а по осі y — значення λ. Цей графік має вигляд кривої, що швидко знижується на початку, а потім вирівнюється.
3. Тепер необхідно знайти "точку зламу" (elbow point), де крива переходить від стрімкого падіння до пологого зниження. Усі компоненти до цієї точки вважаються значущими, решта — шумом. Кількість компонентів до зламу — і є оптимальна кількість факторів.

8. Sequential Mean Test(SMT)

SMT запропонували Yong і Pearce (2013) як більш формалізовану альтернативу паралельному аналізу. Основна ідея полягає в послідовному (sequential) порівнянні кожного eigenvalue з його очікуваним середнім значенням, отриманим зі випадкових даних, доки не буде виявлено перший незначущий eigenvalue — після цього всі наступні автоматично вважаються шумом [20].

У кожному кроці метод порівнює реальне eigenvalue з середнім eigenvalue з даних без латентної структури (тобто випадкових). Тест виконується послідовно: як тільки реальне eigenvalue не перевищує середнє очікуване з випадкових даних — зупиняємося, бо це вказує на межу інформативності.

Алгоритм SMT:

1. Спочатку розраховується кореляційна матриця R_{real}, як завжди, і з неї обчислюються eigenvalues у спадному порядку $\lambda_1^{real}, \lambda_2^{real}, \dots, \lambda_p^{real}$.
2. Далі створюється велика кількість (зазвичай 1000) симульованих наборів випадкових даних, які не мають латентної структури, але мають ті самі розміри. Для кожного з них обчислюються eigenvalues і для кожної компоненти k визначається середнє $\lambda_1^{sim}, \lambda_2^{sim}, \dots, \lambda_p^{sim}$.
3. Для кожного k порівнюється $\lambda_k^{real} > \lambda_k^{sim}$. Перевіряємо послідовно. Якщо вперше виконується $\lambda_k^{real} \leq \lambda_k^{sim}$ — зупиняємося. Кількість значущих факторів — це кількість $k - 1$.

III. Оцінка параметрів моделі

EM (Expectation-Maximization) [21]- це метод максимальної правдоподібності, який використовується для оцінювання параметрів у підтверджувальному факторному аналізі (CFA). Цей метод дає змогу оцінити невідомі параметри моделі шляхом ітеративного процесу максимізації функції правдоподібності і є одним з найбільш популярних методів оцінювання параметрів у CFA та дозволяє отримати більш точні оцінки параметрів моделі в порівнянні з іншими методами оцінювання.

Розглянемо процедуру оцінки параметрів із застосуванням алгоритму EM на прикладі GPCM моделі розмірності 2:

$p_{ijk} = P\left(X_{ij} = k \middle| \overrightarrow{a_i}, d_{ik}, \overrightarrow{\theta_j}\right) = \frac{exp\left((k-1)\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2}\right) + d_{ik}\right)}{\sum_{l=0}^{k} exp\left((l-1)\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2}\right) + d_{il}\right)}$, де $i = \overline{1,N}; j = \overline{1,M}$

N – кількість завдань;

M – кількість іспитників.

Введемо позначення $\Delta_{ik} = (k-1)\left(a_{i1}\theta_{j1} + a_{i2}\theta_{j2}\right) + d_{ik}$, тоді $p_{ijk} = \frac{e^{\Delta ik}}{\sum_{l=0}^{k} e^{\Delta il}}$

$$\mathcal{L} = \prod_{i=1}^{N}\prod_{j=1}^{M} p_{ijk}^{X_{ij}}\left(1 - p_{ijk}\right)^{\left(1-X_{ij}\right)},\ X_{ij} \in \{0, \dots, k\}$$

$$\mathcal{L}^* = \ln \mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} X_{ij} \ln p_{ijk} + \left(1 - X_{ij}\right)\ln\left(1 - p_{ijk}\right)$$

$$\begin{cases} \dfrac{\partial \mathcal{L}^*}{\partial a_{is}} = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M}\left(X_{ij}(k-1) - \frac{X_{ij}\sum_{l=0}^{k-1}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k-1} e^{\Delta_{ijl}}} + \frac{\sum_{l=0}^{k-1}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k-1} e^{\Delta_{ijl}}} - \frac{\sum_{l=0}^{k}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k} e^{\Delta_{ijl}}}\right)\overrightarrow{\theta_j} = 0,\ s = 1,2; \\ \dfrac{\partial \mathcal{L}^*}{\partial \theta_{js}} = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M}\left(X_{ij}(k-1) - \frac{X_{ij}\sum_{l=0}^{k-1}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k-1} e^{\Delta_{ijl}}} + \frac{\sum_{l=0}^{k-1}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k-1} e^{\Delta_{ijl}}} - \frac{\sum_{l=0}^{k}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k} e^{\Delta_{ijl}}}\right)\overrightarrow{a_i} = 0,\ s = 1,2; \\ \dfrac{\partial \mathcal{L}^*}{\partial d_{ik}} = \displaystyle\sum_{i=1}^{N}\sum_{j=1}^{M}\left(X_{ij}(k-1) - \frac{X_{ij}\sum_{l=0}^{k-1}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k-1} e^{\Delta_{ijl}}} + \frac{\sum_{l=0}^{k-1}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k-1} e^{\Delta_{ijl}}} - \frac{\sum_{l=0}^{k}(l-1)e^{\Delta_{ijl}}}{\sum_{l=0}^{k} e^{\Delta_{ijl}}}\right) = 0 \end{cases}$$

У цій роботі для оцінки параметрів використовується функція mirt у мові програмування R [7]. Якщо бути точним то у mirt присутні методи за допомогою яких ця функція може оцінити латентні параметри.

IV. Перевірка адекватності моделей MIRT

Після того як ми провели попередню оцінку розмірності моделі ми отримали можливі розмірності для наших даних. Також ми, знаючи можливі розмірності, спробували оцінити латентні параметри користуючись переліком моделей. Тепер виникає питання модель якої розмірності описує наші дані найкраще. Для цього існують такі методи оцінки адекватності моделі(model fit). В цій роботі використано такі критерії: AIC,BIC, M2, RMSEA, TLI, CFI. Використання декількох критеріїв надає повну картинку для аналізу. Також не слід забувати особливості кожного з критеріїв.

AIC

AIC(Akaike Information Criterion) [10] — це інформаційний критерій, який дозволяє оцінити відносну якість статистичних моделей. Він не говорить, наскільки добре модель пасує до даних сама по собі, але дозволяє порівнювати кілька моделей: чим менше значення AIC, тим краща модель з погляду балансу між підгонкою і складністю.

Статистика AIC визначається математично за допомогою наступної формули:

$$AIC = -2 \cdot \log \mathcal{L} + 2k$$

- $\log \mathcal{L}$ – логарифм максимальної правдоподібності (тобто наскільки добре модель «пояснює» дані)
- k – кількість параметрів у моделі

Інтерпретувати отримані значення статистики AIC можна наступним чином. Низький AIC – модель добре підходить до даних, немає зайвих параметрів. Високий AIC – або модель погано пасує до даних або забагато параметрів.

BIC

BIC(Bayesian Information Criterion) [11] — це ще один інформаційний критерій для порівняння моделей. Він, як і AIC, враховує якість підгонки та складність моделі, але додатково штрафує складні моделі сильніше, особливо при великій вибірці.

Статистика перевірки адекватності моделі BIC визначається наступним чином:

$$BIC = -2 \cdot \log \mathcal{L} + k \cdot \log N$$

- $\log \mathcal{L}$ – логарифм максимальної правдоподібності (тобто наскільки добре модель «пояснює» дані)
- k – кількість параметрів у моделі
- N – розмір вибірки (sample size)

Щодо індивідуальної інтерпритації значення статистики BIC можна сказати, що для меншого значення модель є вигідніша з погляду, що є більш простою і кращу загальну підгонку до даних. Для більшого значення BIC - навпаки, або модель занадто складна або не дуже добре описує дані.

Тут варто зазначити що AIC та BIC не варто інтерпретувати окремо. Вони працюють в парі.

M2

Алгоритм розрахунку статистики M2) [7]:

1. Створюється таблиця спостережуваних ймовірностей (Observed Probabilities) для відповідей на кожну пару завдань (items).
2. Розраховується очікувана ймовірність (Expected Probabilities) на основі параметрів моделі.
3. Різниця між спостережуваними та очікуваними ймовірностями використовується для побудови матриці залишків.
4. Формується статистика:

$$M2 = (O - E)^T W (O - E)$$

де:

O – вектор спостережуваних ймовірностей;

E – вектор очікуваних ймовірностей;

W – вагова матриця, що враховує коваріаційні зв'язки.

5. Порівнюється це значення з критичним значенням з розподілу χ^2 з відповідною кількістю ступенів свободи df.

Щодо висновків які можна зробити на основі значення статистики $M2$. Саме по собі це значення порівнюється з критичним значенням статистики χ^2. Якщо виявляється що $M2 < \chi^2$ то немає підстав відхиляти гіпотезу про добру відповідність моделі (на певному рівні значущості на якому було взяте χ^2). Якщо $M2 > \chi^2$ то модель не відповідає даним (з ймовірністю помилки на рівні значущості). Разом з M2 статистикою R виводить p-значення. p-значення показує, наскільки сильним є відхилення моделі від даних. Якщо це відхилення дуже малоймовірне (тобто $p < 0{,}05$), тоді модель, швидше за все, неадекватна. Якщо це відхилення очікуване (тобто $p > 0{,}05$), тоді модель добре підходить до даних.

RMSEA

RMSEA [9] оцінює, наскільки добре модель підходить до даних, враховуючи її складність (кількість параметрів). Це абсолютний індекс підгонки, який можна інтерпретувати самостійно, на відміну від AIC/BIC.

Формула статистики RMSEA має наступний вигляд:

$$RMSEA = \sqrt{\frac{\chi^2}{df \cdot N} - \frac{1}{N-1}},$$

де:

χ^2 –χ^2 статистика(chi-square);

df – ступені свободи;

N – розмір вибірки(кількість завдань);

При малій вибірці може бути зависоким, навіть якщо модель добре описує дані. Також ця статистика є нестабільною при $N < 100$, особливо якщо df також низьке.

Інтерпретація $RMSEA$ проходить таким чином:

- якщо $RMSEA = 0$ ідеальна підгонка до даних.
- якщо $RMSEA < 0{,}05$ дуже хороша відповідність моделі.
- якщо $0{,}05 < RMSEA < 0{,}08$ прийнятна відповідність моделі.
- якщо $0{,}08 < RMSEA < 0{,}10$ посередня відповідність моделі.
- якщо $RMSEA > 0{,}10$ погана відповідність – модель слід переглянути.

CFI

CFI [8] порівнює вашу модель із "нульовою моделлю", в якій усі змінні некорельовані. Тобто він показує, наскільки краще ваша модель пояснює дані, порівняно з найгіршим можливим варіантом (коли дані є випадковими і не мають прихованих характеристик).

$$CFI = 1 - \frac{\chi^2_{model} - df_{model}}{\chi^2_{null} - df_{null}},$$

де:

χ^2_{null} – χ^2 статистика для випадкових значень;

df_{null} – ступені свободи для випадкових значень;

χ^2_{model} – χ^2 статистика для моделі;

df_{model} – ступені свободи для моделі.

Якщо $CFI = 1$ це вказує на ідеальну підгонку. Якщо ж $CFI < 0{,}9$ є сенс вибрати іншу модель.

TLI

TLI[8], як і CFI, — це індекс порівняльної підгонки, який оцінює, наскільки добре ваша модель пояснює дані порівняно з нульовою моделлю. Але TLI має ще один фокус — він штрафує складні моделі, тобто бере до уваги ефективність підгонки з урахуванням складності.

$$TLI = \frac{\chi^2_{null}/df_{null} - \chi^2_{model}/df_{model}}{\chi^2_{null}/df_{null} - 1},$$

де:

χ^2_{null} – χ^2 статистика для випадкових значень;

df_{null} – ступені свободи для випадкових значень;

χ^2_{model} – χ^2 статистика для моделі;

df_{model} – ступені свободи для моделі.

Якщо $TLI = 1$ це вказує на ідеальну підгонку. Значення TLI може бути нижчим ніж CFI, особливо для складних моделей.

V. Аналіз контрольної роботи і вибір адекватної моделі MIRT

1. Розроблена у магістерській дисертації методика була застосована до аналізу контрольної роботи з дисципліни «Математичний аналіз-1» для бакалаврів 1-го курсу РТФ. Об'єм вибірки склав – 79 студентів. Кількість завдань-16.

2. Для обробки даних було використано мову R. Зокрема, пакети EFA.dimension та mirt [22].

3. Перед використанням результатів контрольної роботи їх було переведено в цілу шкалу. Максимальний бал за кожне завдання в наших даних був 0,63, але кількість категорій (сходинок) в кожному завданні була різна кількість. На рисунках 1-2 показано можливу кількість сходинок і правила перетворення даних.

	A	B	C	D
1	Мінімум за завдання	0		
2	Максимум за завдання	0,6300000		
3	Категорій.12	0,0525000	Шкала.0.063	Шкала.0.12
4	0	0,000	0,00	0
5	1	0,053	0,05	1
6	2	0,105	0,10	2
7	3	0,158	0,15	3
8	4	0,210	0,21	4
9	5	0,263	0,26	5
10	6	0,315	0,31	6
11	7	0,368	0,36	7
12	8	0,420	0,42	8
13	9	0,473	0,47	9
14	10	0,525	0,52	10
15	11	0,578	0,57	11
16	12	0,630	0,63	12
17				
18	Категорій.4	0,1575000	Шкала.0.063	Шкала.0.4
19	0	0,000	0,00	0
20	1	0,158	0,15	1
21	2	0,315	0,31	2
22	3	0,473	0,47	3
23	4	0,630	0,63	4

Рисунок 1. *Закон перетворення даних в цілу шкалу (Категорій.12, Категорій.4).*

	A	B	C	D
25	Категорій.5	0,1260000	Шкала.0.063	Шкала.0.5
26	0	0,000	0,00	0
27	1	0,126	0,12	1
28	2	0,252	0,25	2
29	3	0,378	0,37	3
30	4	0,504	0,50	4
31	5	0,630	0,63	5
32				
33	Категорій.1	0,6300000	Шкала.0.063	Шкала.0.1
34	0	0,000	0,00	0
35	1	0,630	0,63	1
36				
37	Категорій.2	0,3150000	Шкала.0.063	Шкала.0.2
38	0	0,000	0,00	0
39	1	0,315	0,31	1
40	2	0,630	0,63	2
41				
42	Категорій.6	0,1050000	Шкала.0.063	Шкала.0.6
43	0	0,000	0,00	0
44	1	0,105	0,10	1
45	2	0,210	0,21	2
46	3	0,315	0,31	3
47	4	0,420	0,42	4
48	5	0,525	0,52	5
49	6	0,630	0,63	6

Рисунок 2. *Закон перетворення даних в цілу шкалу(Категорій.5, Категорій.1, Категорій.2, Категорій.6).*

4. Після перетворення отримали відповідні цілі значення деякі з них наведені на рисунку 3.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	12	4	0	1	1	1	0	5	1	1	2	1	6	0	1	2
2	7	4	1	1	1	1	1	2	0	0	2	1	1	0	2	2
3	12	4	1	1	1	1	1	5	1	1	2	1	6	1	0	2
4	9	4	1	1	1	1	0	5	1	1	2	1	2	1	2	0
5	12	2	1	1	0	0	1	4	1	1	2	1	6	2	2	2
6	10	4	1	1	1	0	0	5	1	1	2	1	6	2	1	2
7	10	3	0	1	1	0	0	1	1	1	2	0	6	2	2	2
8	5	4	1	1	1	0	0	1	0	1	1	1	4	0	1	1
9	12	4	1	1	1	0	0	5	0	1	2	1	1	0	1	2
10	12	4	1	1	0	0	0	2	1	1	1	1	6	2	1	2
11	12	2	1	1	1	0	0	1	1	1	1	0	6	1	1	1
12	11	4	1	1	1	0	0	5	1	1	2	1	1	0	2	2
13	10	2	1	1	1	0	0	5	1	1	2	1	4	2	1	2
14	8	4	1	1	1	0	0	5	1	1	1	0	1	1	0	1
15	12	4	0	1	1	0	0	4	1	1	2	1	4	2	2	2
16	11	4	1	0	1	1	1	5	1	1	2	0	6	1	2	2
17	8	4	1	1	1	1	0	2	1	1	2	1	2	0	2	2
18	10	4	1	1	1	0	0	5	0	0	2	1	6	1	2	2
19	12	1	1	0	0	0	0	2	1	1	1	0	0	0	2	2
20	9	4	1	1	1	1	1	3	1	1	2	1	0	1	1	2
21	12	4	1	1	0	0	0	0	0	1	2	1	0	1	2	0
22	7	0	1	1	1	0	0	0	0	0	2	1	5	0	0	0
23	12	4	1	1	1	0	0	5	1	1	2	1	6	2	2	2
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Рисунок 3. *Перетворені цілі дані. Перші 23 спостереження.*

5. Після цього було проведено EFA за допомогою алгоритмів CD, EKC, HULL, MAP, NEVALSGT1, RAWPAR, SESCREE, SMT, які реалізовано у функції DIMTEST пакету EFA.dimension. Результати наведено у таблиці 1.

Таблиця 1. *Результати EFA.*

Метод	Розмірність
CD	2
EMPKC	2
HULL	2
MAP	1
NEVALSGT1	7
RAWPAR	2
SESCREE	3
SMT	1

Скрипт виклику методів дослідницького аналізу наведено на рисунку 4.

```
1  library(EFA.dimensions)
2  data<-read.table("conv/ДКР-2,3_цілі.txt",header=FALSE, sep="\t", stringsAsFactors=FALSE)
3  t <- c('CD', 'EMPKC', 'HULL', 'MAP', 'NEVALSGT1', 'RAWPAR', 'SESCREE', 'SMT')
4  DIMTESTS(data, tests = t, corkind='pearson', display=2)
5
```

Рисунок 4. *Виклик методів EFA.*

6. Результат NEVALSGT1 було прийнято як похибку. Було проведено оцінку латентних параметрів для розмірностей S=1,2,3. Результати наведені далі в таблицях 2-10.

Таблиця 2. *Результати оцінювання GPCM моделі S=1.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	0.253	0.177	-0.058	0.682	0.551	0.764	0.394	0.107	0.026	0.447	0.629	0.773	0.463	0.542	0.387	0.622
ak0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ak1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ak2	2	2						2			2		2	2	2	2
ak3	3	3						3					3			
ak4	4	4						4					4			
ak5	5	0						5					5			
ak6	6	-1.999											6			
ak7	7	0.451														
ak8	8	-0.826														
ak9	9	2.093														
d0	0	0.177	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d1	0.26	0	2.185	1.979	1.457	-1.365	-1.189	-0.588	1.625	1.598	2.305	0.805	-0.056	-0.153	-0.635	0.223
d2	1.861	1						-0.117			2.842		0.185	-0.037	-0.213	1.362
d3	3.042	2						-2.504					-0.551			
d4	2.559	3						-0.311					-0.131			
d5	3.147	4						1.116					-1.266			
d6	2.819	0											-0.173			
d7	3.339	-1.999														
d8	2.864	0.451														
d9	3.608	-0.826														

Таблиця 3. *Результати оцінювання GPCM моделі S=2.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	-0.235	-0.188	0.04	-0.291	-0.147	-1.063	-1.233	-0.304	0.108	0	-0.522	-0.493	-0.336	-0.368	-0.568	-0.993
a2	-0.462	0.03	0.158	-0.515	-0.899	0.115	0.812	0.227	-0.311	-1.223	-0.278	-0.396	-0.231	-0.481	0.14	0
ak0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ak1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ak2	2	2						2			2		2	2	2	2
ak3	3	3						3					3			
ak4	4	4						4					4			
ak5	5							5					5			
ak6	6												6			
ak7	7															
ak8	8															
ak9	9															
d0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d1	0.863	-1.983	2.194	1.943	1.59	-1.49	-1.594	-0.336	1.661	1.943	2.268	0.777	-0.107	-0.13	-0.576	0.552
d2	2.902	0.473						0.295			2.803		0.109	-0.039	-0.24	1.672
d3	4.405	-0.8						-2.025					-0.632			
d4	4.147	2.117						0.146					-0.195			
d5	4.871							1.46					-1.293			
d6	4.596												-0.141			
d7	5.082															
d8	4.483															
d9	5.002															

Таблиця 4. *Результати оцінювання GPCM моделі S=3.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	-0.246	-0.172	0.095	2.386	-0.084	-1.097	-1.303	-0.337	-0.061	-0.283	-0.494	-0.353	-0.332	-0.365	-0.541	-1.082
a2	-0.451	0.05	0.144	-4.065	-0.931	0.125	0.866	0.244	-0.777	-2.235	-0.272	-0.413	-0.217	-0.434	0.17	0
a3	-0.029	-0.09	-0.326	-18.834	-0.397	-0.774	-0.099	0.077	2.288	1.396	-0.21	-1.073	-0.095	-0.036	0	0
ak0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ak1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ak2	2	2						2			2		2	2	2	2
ak3	3	3						3					3			
ak4	4	4						4					4			
ak5	5							5					5			
ak6	6												6			
ak7	7															
ak8	8															
ak9	9															
d0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d1	0.812	-1.971	2.246	21.282	1.641	-1.611	-1.631	-0.29	2.861	2.922	2.285	0.933	-0.106	-0.145	-0.581	0.634
d2	2.835	0.495						0.37			2.823		0.109	-0.039	-0.235	1.753
d3	4.325	-0.774						-1.936					-0.633			
d4	4.058	2.145						0.23					-0.196			
d5	4.778							1.524					-1.293			
d6	4.503												-0.137			
d7	4.992															
d8	4.397															
d9	4.919															

Таблиця 5. *Результати оцінювання GRM моделі S=1.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	0,989	0,406	-0,068	0,719	0,587	0,842	0,424	0,327	-0,015	0,455	0,741	0,819	1,514	0,803	0,563	0,995
d1	4,773	2,236	2,185	1,996	1,468	-1,393	-1,195	1,642	1,625	1,6	3,172	0,815	1,78	0,616	0,299	1,669
d2	4,043	2,1						1,081			0,436		0,79	-0,687	-0,644	0,571
d3	2,861	1,1						0,422					-0,127			
d4	1,529	0,904						0,368					-0,565			
d5	1,051	0,406						-0,096					-1,357			
d6	0,385							0,327					-1,745			
d7	-0,053															
d8	-0,799															
d9	-1,371															

Таблиця 6. *Результати оцінювання GRM моделі S=2.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	-0.966	-0.559	-0.04	-0.524	-0.356	-1.31	-1.038	-0.677	0.267	-0.053	-0.699	-0.707	-1.289	-0.625	-0.802	-1.21
a2	-1.436	0.326	0.255	-0.291	-0.765	0.404	0.857	0.618	-0.56	-1.533	-0.231	-0.184	-0.58	-0.652	0.324	0
d1	5.727	2.329	2.209	1.945	1.56	-1.637	-1.533	1.887	1.747	2.131	3.164	0.796	1.715	0.625	0.331	1.786
d2	4.852	2.189						1.23			0.435		0.763	-0.71	-0.688	0.613
d3	3.515	1.157						0.443					-0.12			
d4	1.889	0.953						0.381					-0.543			
d5	1.264							-0.143					-1.311			
d6	0.409												-1.691			
d7	-0.126															
d8	-1.025															
d9	-1.712															

Таблиця 7. *Результати оцінювання GRM моделі S=3.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	-0.71	-0.306	0.16	5.55	-0.032	-1.081	-1.462	-0.893	-0.22	-0.426	-0.569	-0.32	-1.188	-0.571	-0.798	-1.695
a2	-1.015	0.297	0.316	4.791	-0.586	0.621	0.999	0.44	-1.769	-2.592	-0.167	0.144	-0.477	-0.597	0.303	0
a3	-0.708	-0.399	-0.317	-24.487	-0.79	-0.706	0.397	0.254	1.476	0.01	-0.53	-1.35	-0.825	-0.415	0	0
d1	5.266	2.327	2.269	26.444	1.615	-1.677	-1.76	1.92	2.824	2.933	3.216	0.944	1.778	0.632	0.328	2.072
d2	4.462	2.19						1.212			0.443		0.798	-0.71	-0.686	0.719
d3	3.219	1.159						0.389					-0.115			
d4	1.726	0.954						0.326					-0.556			
d5	1.154							-0.195					-1.356			
d6	0.378												-1.749			
d7	-0.107															
d8	-0.931															
d9	-1.567															

Таблиця 8. *Результати оцінювання NRM моделі S=1.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	0.393	0.313	-0.383	0.016	0.246	0.433	0.457	0.226	0.085	0.241	0.47	0.365	4.103	0.63	0.463	0.689
ak0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ak1	-8.859	1.611	1	1	1	1	1	6.449	1	1	1.668	1	2.141	0.985	1.312	0.552
ak2	8.255	3.922						7.194			2		0.186	2	2	2
ak3	3.788	7.388						1.413					4.207			
ak4	3.349	4						5.256					5.431			
ak5	6.283							5					-0.702			
ak6	4.346												6			
ak7	9.004															
ak8	7.939															
ak9	9															
d0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d1	-6.552	-1.704	2.242	1.822	1.388	-1.27	-1.203	-0.245	1.627	1.554	2.272	0.734	2.495	-0.118	-0.609	0.147
d2	3.26	0.876						0.136			2.786		0.32	-0.041	-0.194	1.346
d3	3.94	-1.017						-2.34					0.537			
d4	3.208	2.502						0.068					-1.455			
d5	4.256							1.513					-5.071			
d6	3.625												-3.155			
d7	4.222															
d8	3.992															
d9	4.683															

Таблиця 9. *Результати оцінювання NRM моделі S=2.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	-1.501	0.007	-0.319	0.057	0.021	-1.084	-1.258	-4.782	0.001	-0.148	-0.575	-0.39	-0.299	-0.416	-0.555	-1.679
a2	-3.144	0.04	0.639	-0.697	-0.742	-0.223	0.49	2.626	0.259	-0.55	-0.157	-0.335	-0.282	-0.481	0.121	0
ak0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ak1	-2.247	0.391	1	1	1	1	1	0.819	1	1	1.367	1	0.699	1.733	1.731	-0.249
ak2	4.837	9.561						5.42			2		-0.188	2	2	2
ak3	1.336	-34.984						6.209					4.508			
ak4	3.385	4						5.251					2.485			
ak5	3.157							5					-2.557			
ak6	7.111												6			
ak7	9.345															
ak8	4.215															
ak9	9															
d0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d1	-18.939	-2.078	2.377	1.993	1.525	-1.501	-1.521	4.414	1.646	1.638	2.411	0.757	-0.249	-0.102	-0.591	-0.437
d2	17.195	0.291						17.518			2.928		-0.304	0.083	-0.146	1.597
d3	10.401	-1.974						7.375					-1.257			
d4	16.146	1.952						18.101					-0.41			
d5	16.228							20.055					-2.315			
d6	17.337												-0.691			
d7	13.913															
d8	17.499															
d9	15.525															

Таблиця 10. *Результати оцінювання NRM моделі S=3.*

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
a1	-2.061	0.031	-0.259	3.658	-0.058	-1.128	-1.249	-4.34	-0.007	-0.382	-0.527	-0.529	-0.395	-0.47	-0.533	-1.327
a2	-3.441	0.039	0.802	-15.87	-0.711	0.04	0.756	3.409	0.065	-0.835	-0.122	-0.281	-0.268	-0.426	0.218	0
a3	-0.172	-0.095	-0.591	-29.797	-0.54	-0.594	-0.03	0.678	0.947	0.467	-0.282	-1.773	-0.024	0.057	0	0
ak0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ak1	0.532	-56.782	1	1	1	1	1	1.7	1	1	1.032	1	1.072	1.72	1.507	-0.255
ak2	4.346	-6.247						5.37			2		-0.06	2	2	2
ak3	0.612	-36.989						5.766					3.661			
ak4	2.519	4						5.276					2.485			
ak5	3.863							5					-1.483			
ak6	6.391												6			
ak7	9.14															
ak8	5.508															
ak9	9															
d0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
d1	3.845	-10.921	2.621	38.827	1.602	-1.543	-1.578	10.405	1.865	1.819	2.352	1.238	-0.158	-0.089	-0.545	-0.374
d2	17.672	0.072						21.823			2.909		-0.284	0.095	-0.141	1.55
d3	6.695	-5.482						16.122					-1.018			
d4	13.955	1.928						22.096					-0.33			
d5	17.833							24.208					-2.006			
d6	18.824												-0.781			
d7	14.918															
d8	19.18															
d9	15.899															

Скрипт виклику функції для оцінки параметрів для моделі GPCM розмірності 1 наведено на рисунку 5.

```
5  library(mirt)
6  mod1_gpcm<-mirt(data=data,model=1,itemtype = c("gpcm"), technical = list(NCYCLES=100000))
7  coef(mod1_gpcm)
8
```

Рисунок 5. *Виклик оцінки коефіцієнтів для моделі GPCM розмірність 1.*

7. Після цього було проведено оцінку адакватності моделей. Результати наведені в таблиці 11.

Таблиця 11. *Результати перевірки адекватності моделей.*

	AIC	BIC	M2	RMSEA	CFI	TLI
gpcm1	2286,332	2419,022	31,5653	0,06145317	0	-1,446716
graded1	2285,039	2417,729	26,5008	0,06266108	0	-1,543845
nominal1	2283,802	2473,358	23,63699	0	1	1,64714
gpcm2	2286,812	2455,043	19,62494	0	1	1,302478
graded2	2287,175	2455,406	20,30307	0	1	1,472419
nominal2	2261,388	2486,486	20,71856	0,05512665	0,1591264	-0,9688748
gpcm3	2291,253	2492,656	23,36831	0	1	2,731539
graded3	2290,479	2491,882	27,81973	0	1	3,309482
nominal3	2251,851	2510,121	28,3454	0	1	3,268831

Як бачимо, більшість критеріїв вказує на модель GPCM розмірності 2.

Виклик статистики M2, що включає в себе значення статистик RMSEA, CFI, TLI, а також AIC і BIC довелося проводити окремо. Скрипт викликів наведено на рисунку 6.

```
5  library(mirt)
6  mod1_gpcm<-mirt(data=data,model=1,itemtype = c("gpcm"), technical = list(NCYCLES=100000))
7  M2(mod1_gpcm)
8  extract.mirt(mod1_gpcm, 'AIC')
9  extract.mirt(mod1_gpcm, 'BIC')
10
```

Рисунок 6. *Виклик статистик M2, RMSEA, CFI, TLI, AIC, BIC.*

ВИСНОВКИ

1. Досліджено основні багатовимірні (MIRT) моделі, які можуть бути застосовані для аналізу тестів з вищої математики. А саме, GRM, GPCM, NRM.
2. Розширено набір методів дослідницького факторного аналізу, які дозволяють зробити первинний вибір розмірності моделі, а саме, застосовано : CD, EKC, HULL, MAP, NEVALSGT1, RAWPAR, SESCREE, SMT.
3. На основі відібраних методів дослідницького факторного аналізу проведено первинне визначення розмірності даних.
4. Проведено підтверджувальний факторний аналіз на основі алгоритму EM шляхом оцінки параметрів моделей.
5. Знайдено значення відповідних латентних характеристик іспитників і завдань згідно моделей MIRT.
6. Проведено перевірку адекватності моделей MIRT на підставі критеріїв: AIC, BIC, M2, RMSEA, CFI, TLI.
7. Показано, що для політомічних даних «найкращою» є модель GPCM розмірності 2.
8. На основі відібраної моделі доречно проводити аналіз тестів з вищої математики, а саме, видаляти надто прості або складні завдання, завдання з низькою розрізняльною здатністю.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Crocker L., Algina J. Introduction to Classical and Modern Test Theory. — Belmont, CA:Wadsworth — 2006. — 527 p.
2. W. Linden, R. Hambleton Handbook of modern item response theory — New York: Springer — 2018. — 1500 p.
3. Reckase M. D. Multidimensional item response theory — New York: Springer — 2009 — 354 p.
4. Mahmood Ul H., Frank M. Discrimination with unidimensional and multidimensional item response theory models for educational data // Communications in Statistics - Simulation and Computation, — 2019 — p. 1–21.
5. Auerswald, A., and Moshagen, M. How to Determine the Number of Factors to Retain in Exploratory Factor Analysis: A Comparison of Extraction Methods Under Realistic Conditions. Psychological Methods. Advance online publication. — 2019 — p. 1-24.
6. Chalmers, R. Multidimensional Item Response Theory Package for the R Environment. Journal of Statistical Software, Volume 48, Issue 6 — 2012 — p. 1-25.
7. R. Philip Chalmers. Mirt: A Multidimensional Item Response Theory Package for the R Environment.
8. Yu, C. Y. Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. Unpublished doctoral dissertation, University of California, Los Angeles, 2002. https://escholarship.org/content/qt7k49h7pm/qt7k49h7pm.pdf
9. Maydeu-Olivares, A., & Joe, H. Limited information goodness of-fit testing in multidimensional contingency tables. Psychometrika, 71(4), — 2006 — p. 713-732.
10. Muthén, B., & Muthén, L. How to use a Monte Carlo study to decide on sample size and determine power. Structural Equation Modeling, 9(4), — 2002 — p. 599-620
11. Muthén, B., & Muthén, L. K. Mplus user's guide. Los Angeles, CA: Muthén & Muthén, — 2012.
12. Ruscio, J., & Roche, B. Determining the number of factors to retain in an exploratory factor analysis using comparison data., Psychological Assessment, 24(2), — 2012 — p. 282–292.
13. Kaiser, H. F. The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20(1), 1960, p. 141–151.
14. Braeken, J., & van Assen, M. A. An empirical Kaiser criterion. Psychological Methods, 22 — 2017 — p. 450–466.

15. Lorenzo-Seva, U. A factor simplicity index. Behavior Research Methods, 38(4) — 2006 —p. 630–637.
16. Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. The Hull method for selecting the number of common factors. Multivariate Behavioral Research, 46(2) — 2011 — p. 340–364.
17. Velicer, W. F. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41(3) — 1976 — p. 321–327.
18. Horn, J. L. A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2) — 1965 — p. 179–185.
19. Cattell, R. B. The Scree Test For The Number Of Factors. Multivariate Behavioral Research, 1(2) — 1966 — p. 245–276.
20. Yong, A. G., & Pearce, S. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. Tutorials in Quantitative Methods for Psychology, 9(2) — 2013 — p.79–94
21. Dempster A.P, Laird N.M., Rubin D.B. Maximum Likelihood From Incomplete Data via the EM Algorithm., Journal of the Royal Statistical Society B, 39(1) — 1977 — p. 1–38.
22. The Comprehensive R Archive Network. http://cran.us.r-project.org