МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

МАТЕМАТИЧНИЙ ФАКУЛЬТЕТ
Кафедра загальної математики

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему: **«ПОРІВНЯЛЬНИЙ АНАЛІЗ ОКРЕМИХ АВТОМАТИЗОВАНИХ СИСТЕМ ОБРОБКИ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ»**

Виконала: студентка 2 курсу, групи 8.1118

спеціальності 111 математика
(шифр і назва спеціальності)

освітньої програми математика
(назва освітньої програми)

А. А. Шека
(ініціали та прізвище)

Керівник завідувач кафедри загальної математики, доцент, к.ф.-м.н., доцент Зіновєєв І.В.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Рецензент доцент кафедри фундаментальної математики, доцент, к.ф.-м.н. Клименко М.І.
(посада, вчене звання, науковий ступінь, прізвище та ініціали)

Запоріжжя – 2020

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ЗАПОРІЗЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет математичний

Кафедра загальної математики

Рівень вищої освіти магістр

Спеціальність 111 математика
(шифр і назва)

Освітня програма математика

ЗАТВЕРДЖУЮ

Завідувач кафедри загальної математики, доцент, к.ф.-м.н.

_____ Зіновєєв І. В.
(підпис)

« 30 » травня 2019 р.

З А В Д А Н Н Я

НА КВАЛІФІКАЦІЙНУ РОБОТУ СТУДЕНТЦІ

Шеці Анні Анатоліївні
(прізвище, ім'я та по-батькові)

1. Тема роботи Порівняльний аналіз окремих автоматизованих систем обробки результатів тестування

керівник роботи Зіновєєв Ігор Валерійович, к.ф.-м.н, доцент
(прізвище, ім'я та по-батькові, науковий ступінь, вчене звання)

затверджені наказом ЗНУ від « 29 » травня 2019 року № 811-с

2. Строк подання студентом роботи 26 грудня 2019 року

3. Вихідні дані до роботи 1. Постановка задачі.
2. Перелік літератури.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)
1. Тестологія – наука про створення якісних та науково обґрунтованих вимірювальних діагностичних тестів
2. Теоретичні положення обробок результатів тестування та автоматизовані системи на їх основі
3. Обробка результатів тестування автоматизованими системами Iteman, Ministep та Їх порівняльний аналіз

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень) ілюстрації до тексту, презентація

6. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв

7. Дата видачі завдання 30.05.2019

К А Л Е Н Д А Р Н И Й П Л А Н

№	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1.	Розробка плану роботи.	13.06.2019	
2.	Збір вихідних даних.	08.08.2019	
3.	Обробка методичних та теоретичних	20.09.2019	
	джерел.		
4.	Розробка першого та другого розділу.	07.10.2019	
5.	Розробка третього розділу.	20.11.2019	
6.	Оформлення та нормоконтроль	15.12.2019	
	кваліфікаційної роботи.		
7.	Захист кваліфікаційної роботи.	16.01.2020	

Студентка _____ А. А. Шека
(підпис) (ініціали та прізвище)

Керівник роботи _____ І. В. Зіновєєв
(підпис) (ініціали та прізвище)

Нормоконтроль пройдено

Нормоконтролер _____ О. Г. Спиця
(підпис) (ініціали та прізвище)

РЕФЕРАТ

Кваліфікаційна робота магістра: «Порівняльний аналіз окремих автоматизованих систем обробки результатів тестування»: 110 с., 55 рис., 12 табл., 34 джерел, 1 додаток.

ВАЛІДНІСТЬ ТЕСТУ, ДИСКРИМІНАТИВНІСТЬ, ДИСТРАКТОР, НАДІЙНІСТЬ ТЕСТУ, ПЕДАГОГІЧНИЙ ТЕСТ, РЕПРЕЗЕНТАТИВНІСТЬ, ТЕСТ, ТЕСТОЛОГІЯ, ТЕСТУВАННЯ, CRT, ITEMAN, IRT, MINISTEP, MYTESTXPRO.

Об’єкт дослідження – основи теорій CRT та IRT, автоматизовані системи обробки результатів тестування на положення цих теорій, методика дослідження тестів, педагогічні тести.

Мета роботи: ознайомитись з основними положеннями CRT та IRT теоріями обробки результатів тестування та дослідити якість пробного тесту ЗНО з математики, що проводилось на базі навчального закладу ЗНУ у 2019 році.

Методи дослідження – аналітичний, синтез-метод, вимірювання, порівняння, аналіз.

У кваліфікаційній роботі розглядаються необхідні теоретичні відомості з тестології, її розвиток та критерії щодо якості тесту. Розглянуто положення CRT та IRT теорій, автоматизовані системи на їх основі. Зроблено порівняльний аналіз систем.

На базі дослідженого матеріалу було проведено порівняльний аналіз автоматизованих систем Iteman, Ministep обробки результатів тестування на основі CRT та IRT теорій. Досліджено якість пробного тесту ЗНО з математики, що проходив у навчальному закладі ЗНУ у 2019 році за допомогою програм MS Excell, Iteman, Ministep. Результати порівняльного аналізу роботи можуть бути використанні для детального дослідження якості педагогічних тестів.

SUMMARY

Master's Qualification Thesis «Comparative Analysis of Certain Automated Systems for Processing Test Results»: 110 pages, 55 figures, 12 tables, 34 references,1 supplements.

TEST VALIDITY, DISCRIMINATORY POWER, DISTRACTOR, TEST RELIABILITY, PEDAGOGICAL TEST, REPRESENTATION, TEST, TESTOLOGY, TESTING, CRT, ITEMAN, IRT, MINISTEP, MYTESTXPRO.

The object of the study is the basics of CRT and IRT theories, automated systems for the processing the results of the testing on the principles of these theories, the methodology of the investigation of the tests, the pedagogical tests.

The aim of the study is to get acquainted with the basic principles of CRT and IRT theories of the processing of the test results and to investigate the quality of the Exam Focus Complex in mathematics, held at the Zaporozhe National university in 2019.

The methods of research are analytical, synthesis-method, measurement, comparison and analysis.

The qualification work deals with the necessary theoretical information of the testology, its development, and criteria for the quality of the test. The main principles of CRT and IRT theories, automated systems based on them have been considered. The comparative analysis of systems has been made.

The comparative analysis of the automated systems Iteman, Ministep for the processing the results of the testing based on CRT and IRT theories has been the studied material. The quality of the Exam Focus Complex in mathematics, which was held at ZNU in 2019 using MS Excell, Iteman, Ministep has been investigated. The results of the comparative analysis of the work can be used for a detailed study of the quality of pedagogical tests.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ВНЗ – вищий навчальний заклад

ЗНО – зовнішнє незалежне оцінювання

ASC – Assessment Systems Corporation

CRT – Classical Response Theory

ETS – Educational Testing Service

GUI – graphical user interface

IRT – Item Response Theory

JML – joint maximum likelihood procedure

NAEP – National Assessment of Educational Progress

RM – Rasch Measurement

RUMM – Rasch Unidimensional Measurement Model

ВСТУП

Актуальність теми. Тести, що використовуються на сьогодні майже у всіх сферах діяльності, повинні відповідати таким характеристикам, як надійність, валідність та ефективність. Для виявлення недоліків тестів та їх покращення використовують різноманітні теорії та автоматизовані системи, що засновані на їх положеннях. Аналіз ефективності цих теорій та систем є задачею актуальною.

Мета та завдання дослідження: ознайомитись з основними положеннями CRT та IRT теоріями обробки результатів тестування та дослідити якість пробного ЗНО з математики, що проводилось на базі навчального закладу ЗНУ у 2019 році.

Об'єкт дослідження –теорії CRT та IRT, автоматизовані системи обробки результатів тестування на положення цих теорій, методика дослідження тестів, педагогічні тести.

Методи дослідження. У роботі використовуються такі методи, як аналітичний, синтез-метод, вимірювання, порівняння, аналіз.

Практичне значення одержаних результатів. Результати роботи можуть бути використанні для детального дослідження якості педагогічних тестів, покращення системи навчання, професійної підготовки фахівців.

Структура й обсяг кваліфікаційної роботи. Робота складається з трьох розділів. У першому розділі наведено історичні аспекти створення та розвитку тестології як науки у світі і в Україні, а також основні її поняття та критерії щодо якості тесту. У другому розділі наведено теоретичні основи підходів до обробки результатів тестування – Classic Response Theory і Item Response Theory. Також розглянуті автоматизовані системи, засновані на положення цих підходів. У третьому розділі наведено результати дослідження якості пробного ЗНО з математики, за допомогою таких програм, як MS Excell, Iteman, Ministep, створено порівняльний аналіз систем.

1 ТЕСТОЛОГІЯ – НАУКА ПРО СТВОРЕННЯ ЯКІСНИХ ТА НАУКОВО ОБҐРУНТОВАНИХ ВИМІРЮВАЛЬНИХ ДІАГНОСТИЧНИХ ТЕСТІВ

1.1 Історія створення та розвиток науки тестології

Перші відомості про перевірку знань і здібностей окремих людей за спеціальними завданнями відносять до IV-III ст. до н.е. Дослідження Теофана (372-283 рр. до н.е.) «Характери», присвячену психологічній діагностиці, справедливо можна вважати першоджерелом педагогічної діагностики. Початковий період її формування пов’язаний з іменами Аристотеля, Платона, Галля і Лафатера [1].

У стародавньому Вавилоні застосовувалися різноманітні випробовування, що дозволяли перевіряти здібності людей і отримані ними знання щодо оволодіння професією писаря. У Китаї (III ст. до н.е. - III ст. н.е.) проводили спеціальні іспити з метою відбору претендентів на посаду урядових чиновників. У літературних джерелах є немало свідчень щодо застосування різноманітних видів випробувань у стародавній Греції, Спарті та інших державах [2].

Уперше педагогічне тестування (від англ. «testing» – випробування, проба, іспит) використав Реверенд Джордж Фішер для перевірки рівня знань студентів за допомогою оригінальних спеціальних книг («scale books»), що з’явилися у 1864 р. (Великобританія) [2]. Ці книги можна вважати першими зразками шкільних тестів успішності, але теоретичні основи тестування були розроблені пізніше, у 1883 році, також англійським психологом Френсісом Гальтоном у його роботі «Дослідження людських здібностей та їх розвиток» [3]. Досліджуючи індивідуальні відмінності, Френсіс Гальтон використовував цілий набір методик: на визначення зорової, слуховий, тактильної чутливості, на мускульну силу, час реакції і ін. Під час Міжнародної виставки медичного обладнання, засобів і методів охорони здоров’я в Лондоні у 1884 році була влаштована лабораторія, де відвідувачі у віці від 5 до 80 років могли перевірити свої фізичні здібності, фізіологічні

можливості організму і психічні властивості по 17 показникам: ріст, вага, життєва ємкість легень, сила удару, розрізнення кольорів, гострота зору і ін.

Було обстежено понад 9000 чоловік. Важливим внеском Гальтона в розвиток теорії тестів було визначення трьох основних принципів:

а) застосування серії однакових випробувань до великої кількості випробовуваних;

б) статистична обробка результатів;

в) виділення еталонів оцінки.

Ці принципи використовуються і до сьогодення – на основі проведення серій випробувань отримують різного виду норми для оцінки результатів тестування, всі сучасні тести побудовані на основі статистичної теорії вимірювань, а ідея еталона оцінки лежить в основі визначення тестів як стандартизованого інструменту.

Френсіс Гальтон називав випробовування, що проводились у його лабораторії, – розумовими тестами.

Експерименти Гальтона з тестування фізичних можливостей організму та психологічних властивостей людини передвизначили перший відхід від тисячолітньої практики випробувань і перевірок, заснованої на інтуїції, і перехід до фундаментального дослідження проблеми. Отже, вже до кінця XIX століття експеримент почали розглядати як обов’язкову умову подальшого розвитку тестування.

Першим дослідником, який застосував у психологічному експерименті інтелектуальний тест, був Джеймс Мак-Кін Кеттел [1]. Цей термін після статті Кеттела «Інтелектуальні тести та вимірювання», що опублікована у 1890 році, набув широкого розповсюдження. У цій статті висловлювалася думка про те, що наукова і практична цінність тестів зросте, якщо умови їх проведення будуть одноманітними. Так, вперше було проголошено необхідність стандартизації тестів для можливості порівняння їх результатів, отриманих різними дослідниками на різних випробовуваних.

Він поставив завдання описати образ повноцінної особистості за допомогою можливо меншого числа експериментів. З цією метою він

запропонував декільком лабораторіям зробити в однакових умовах 10 основних експериментів (вимір сили рук за допомогою динамометра, швидкості реакції на звук, швидкості асоціації при назві 10 кольорів і т.д.). На цій основі згодом їм були розроблені набори завдань, які він називав «розумовими тестами» [3]. Таких тестів Дж. Кеттеллом було розроблено 50. Однак всі вони дозволяли оцінити елементарні психічні процеси, де індивідуальні відмінності порівняно малі і не зачіпали вищих психічних функцій, що лежать в основі інтелекту.

Розглядаючи тест як випробування з практичною метою, як засіб проведення наукового експерименту з дослідження особистості, Дж. Кеттел сформулював систему вимог до тестування:

а) однаковість умов для всіх випробовуваних;

б) обмеження часу тестування приблизно однією годиною;

в) відсутність глядачів у приміщенні, де проводиться тестування;

г) якість обладнання, наявність інструкцій, однакових для всіх, розуміння випробовуваними поставлених завдань;

д) статистична обробка результатів тестування (розрахунок мінімального, середнього і максимального результатів, середнього арифметичного та стандартного відхилення).

Усі ці ідеї, висунуті Дж. Кеттеллом, у даний час складають основу сучасної класичної тестології.

Засновником тестової діагностики вважається Дж. Кеттел, який започаткував традицію дослідження інтелекту вступників до вищих навчальних закладів саме за допомоги тестів, яка зберігається в американських університетах і донині [4].

Для вивчення ефективності дидактичних прийомів у 1894 році американець Дж. Райс вперше застосував свої таблиці з перевірки знань орфографії, а М. Стоун у 1908 році надрукував перший тест з арифметики [4].

Якісний стрибок у розвитку тестології пов’язаний з діяльністю видатного французького психолога Альфреда Біне [5]. Він може вважатися родоначальником сучасних тестів, призначених для діагностики рівня розвитку інтелекту. У 1904 році А. Біне увійшов до складу комісії зі створення в Парижі

спеціальних шкіл для розумово неповноцінних дітей. Було потрібно відокремити дітей, здатних до навчання, але ледачих і не бажаючих учитися, від дітей з розумовими вадами. А. Біне і Теодор Симон розробили серію завдань для дітей від 3 до 11 років. Спочатку серія складалася з 30 тестів-завдань, розташованих у міру зростання складності таким чином, що ймовірність успішного виконання підвищувалася з хронологічним віком. Рівень складності був знайдений у результаті обстеження 50 нормально розвинених дітей цих вікових груп і незначної кількості дітей з розумовими вадами. Остання група дітей не могла вирішити завдання важче певного рівня складності.

Фактично застосування цього тесту було першою спробою визначити індивідуальні відмінності між дітьми за допомогою вимірювання їх розумового розвитку.

А. Біне і Т.Сімон кілька разів переглядали створену ними шкалу. У 1908 році здійснено нова редакція, в ході якої була поставлено принципово нове завдання – не тільки диференціація дітей з вродженими розумовими вадами і дітей з нормальним розвитком, але і виділення різних рівнів інтелектуального розвитку нормальних дітей.

Важливою зміною було те, що вперше тести групувалися за віковими рівнями, що дозволило визначити норми для дітей різного віку, і вводилося поняття розумового рівня (пізніше заміненого на розумовий вік, а ще пізніше на показник розумового розвитку IQ).

Досить тривалий час тести розвивалися як інструмент індивідуальних вимірювань. Масове використання тестування спонукало перейти від індивідуальних тестів до створення групових.

У 1917-1918 роках у США з’явилися перші групові тести для потреб армії. Основні принципи, використані при складанні цих тестів, були систематизовані і згодом лягли в основу всієї методології групових тестів [2]:

а) принцип обмеження в часі (щоб тільки 5% випробовуваних могли закінчити опрацювання всього тесту), тобто показник розвитку прямо залежить від швидкості виконання завдань випробуваним;

б) принцип деталізованої інструкції як при проведенні, так і під час підрахунку;

в) введено тести з вибірковим методом формування відповіді з зазначенням у разі незнання або вибиранням відповідь навмання;

г) підбір тестів після ретельної статистичної обробки та експериментальної перевірки.

Незважаючи на розпочаті спроби впровадження тестів у навчальні заклади США, підхід до тестування наприкінці XIX століття мав, в основному, теоретичний характер, а тестування все ще залишалося переважно сферою діяльності психологів.

З початку XX століття визначилося і педагогічне спрямування в розвитку тестології [2]. Американець В.А. Макколл розділив тести на педагогічні (Educatinal Test) і психологічні – за визначенням рівня розумового розвитку (Intelligence Test). Основним завданням педагогічних тестів було вимір успішності учнів з тих чи інших шкільних дисциплін за певний період навчання, а також успішності застосування певних методів викладання і організації.

В. Макколл визначив мету використання педагогічних тестів – об'єднання в групи учнів, що засвоюють рівний за обсягом матеріал з однаковою швидкістю.

На початку XX століття у розробці тестів спостерігається розмежування психологічного та педагогічного напрямів [1]. Розробка першого педагогічного тесту належить американському психологу Едуардові Лі Торндайк. Він вважається основоположником педагогічних вимірювань. Перший стандартизований педагогічний тест, який вийшов під керівництвом Е. Торндайк, був тест Стоуна на рішення арифметичних задач, вперше забезпечений «нормами».

У 1915 році американський дослідник Роберт Йеркс створив свою серію тестів, головна відмінність якої – зміна системи підрахунку. Замість вікових часток, запропонованих А. Біне, випробуваний отримує за кожен правильно вирішений тест відоме кількість балів. Це підвищило зручність відносно проведення та підрахунку результату тесту. Кількість отриманих балів

переводилося по прикладеним стандартам в коефіцієнт обдарованості або успішності.

Саме з розвитком тестування у психології та педагогіці почали застосовуватися математичні методи, що мало вплив на розвиток тестології. Цей період характеризується підвищенням інтересу до тестування як засобу оцінювання академічних здібностей. З цього моменту тестування розвивається за двома головними напрямами:

а) створення та використання тестів інтелектуального розвитку;

б) створення та використання педагогічних тестів, призначених для оцінювання академічних здібностей і знань студентів.

Розширення сфер впровадження тестів вимагало розробки загальних вимог до них. Саме С. Пріссей і Ф. Робінсон у праці «Психологія і нова освіта» (1933 р.) виступили з підтримкою стандартизованих тестів, визначивши такі їх особливості:

а) ретельний відбір матеріалу для тестування;

б) чітка ясність і недвозначність вказівок;

в) правильне, з точки розумових норм, формулювання запитань;

г) об’єктивне і по можливості просте оцінювання одержаних результатів.

Поступово тестування із психолого-педагогічної проблеми перейшло в соціальну та ідеологічну сферу. Тести почали піддавати гострій критиці, оскільки за результатами тестування особи, які відносилися до чорної раси, вихідці з менш забезпечених сімей, мали гірші показники, на відміну від представників більш забезпечених прошарків населення.

У колишньому СРСР розвиток і використання діагностичних методів має свою історію [6]. Загалом ця історія не збігається із всесвітньою. Умовно можна виділити три періоди.

Перший – з початку 20-х до середини 30-х років. Цей період визначається поширенням різних тестових методик і наукових пошуків у даній галузі. Але відома постанова ЦК ВКП(б), видана у 1936 р. під назвою «Про педагогічні викривлення у системі наркомпросів», у якій засуджувалася практика використання тестів, зробила свою справу. Тестування було пов’язано з так

званою педологічною наукою і практикою, що припустилася абсолютизації одержаних за допомогою тестів результатів дослідження школярів. Категоричне та загально масштабне засудження педології як псевдонауки супроводжувалося відкиданням позитивних досягнень радянських педагогів і психологів у галузі оцінки знань і вмінь учнів, які в цілому творчо розвивали педагогіку та психологію. Тому проблема тестування в педагогічній і психологічній літературі протягом тривалого часу ігнорувалася, а припинення всіх досліджень, пов'язаних з розробкою та використанням тестів, стало суттєвою перешкодою на шляху подальшого розвитку психодіагностики і дидактодіагностики.

Отже, протягом приблизно двадцяти років в СРСР проблема діагностичних методик навіть не поставала.

Початок другого етапу розвитку методів тестування можна віднести до 60-х років, коли тестування як метод вимірювання знань почали використовувати у військових училищах Міністерства оборони, Міністерства внутрішніх справ та інших спеціалізованих закладах. Цей період відзначається також активізацією досліджень у галузі програмованого навчання. Виходять збірники наукових праць: «Питання алгоритмізації і програмованого навчання» (1973 р.), «Питання теорії і методики програмованого навчання», «Програмоване навчання за кордоном» (1968 р.), роботи Т. Ільїної «Програмоване навчання і школа» (1968 р.), М. Нікандрова «Програмоване навчання та ідеї кібернетики» (1970 р.), Н. Тализіної «Теорія програмованого навчання» (1975 р.) та інших. Нова хвиля повторного використання тестів, починаючи з 60-х років, мала загальносвітовий характер і пов'язана з іменем відомого американського педагога Ральфа Тайлера. За його ініціативою було розроблено програму NAEP, яка суттєво вплинула на впровадження тестування у систему освіти США. Головний принцип Р. Тайлера полягав у тому, що його програма мала на меті не здійснення контролю за місцевими системами шкільної освіти «зверху», вона повинна була викликати зацікавленість у вдосконаленні цього процесу.

Але вже наприкінці 70-х років деякі американські вчені дійшли висновку, що процес оцінки знань не повинен бути обмежений використанням лише стандартизованих тестів. Дослідження, проведені Х. Уеллсом, показали: щоб

допомогти учням у засвоєнні навчального матеріалу, необхідні такі тести, які визначали б щоденні успіхи учня. Такими, на його думку, є тести з відповідних галузей знань (Domain Referenced Tests – DRT).

Дж. Мак і В. Хант підтримують ідею використання тестів типу (DRT) і дають їм назву критеріально-орієнтованих (Criterion-Referenced Tests) [7]. Більшість спеціалістів із проблем тестування вважають, що для учнів тести є безпосередньо тим критерієм, який дає їм змогу краще оцінити себе, з'ясувати мету та методи навчання. При правильному використанні тести допомагають як викладачам, так і учням. А при неправильному – можуть загальмувати процес навчання. Висловлюється також думка про неприпустимість і шкідливість використання з метою навчання одного тесту для цілого класу учнів, тому що для деяких учнів він може виявитися дуже легким, а для інших – важким. Отже, тести мають бути настільки індивідуалізовані, щоб кожен учень одержував тест, який відповідає його рівню засвоєння навчального матеріалу. Викладач готує усі тести на початку курсу, а учні проходять тестування в міру підготовки. Але якщо не буде враховано навчальну функцію тестів, то вони можуть завдати шкоди як учням, так і навчальній програмі.

Частина вчених має іншу думку. До них належить Р. Ебел [8], який вважає, що саме нестандартизовані тести (складені окремими викладачами) частіше містять помилки. Вони обмежені за обсягом матеріалу, а також не включають розділи, важливі для розвитку учня.

В останні роки тестування широко використовується і в європейських країнах, зокрема у Німеччині, Швеції, Норвегії, Великобританії та інших [2]. Так, у Німеччині, на відміну, наприклад, від Англії, впровадження тестування стримувалося, по-перше, політичними обставинами, а по-друге, тим, що там мала перевагу гуманітарна педагогіка. І тільки з середини 50-х рр. у деяких галузях системи освіти почали застосовувати тести. Поштовх поширенню тестування дала Міжнародна конференція в 1967 р. у Берліні, яка стимулювала розвиток тестів, орієнтованих на критерії. Але масове застосування цього методу в системі освіти Німеччини відбулося у середині 80-х рр. і було наслідком

поширення американської системи. На цей період припадає введення при вступі до найпрестижніших університетів обов’язкових тестувань.

Отже, аналіз історичного розвитку тестології засвідчує, що ця галузь набула широкого розвитку у США та розвинених західних країнах, де накопичено великий досвід щодо розробки і практичного застосування тестів у різних сферах діяльності.

У 1983 р. проведенням програми NAEP починає займатися ETS, яка поставила за мету систематизацію робіт з тестування, зокрема стандартизацію тестів, встановлення єдиних правил процедури тестування, а також розробку критеріїв для визначення якості освіти, набутої американськими школярами та студентами [9].

Історія створення та використання технологій об’єктивного контролю налічує більше ніж 140 років, коли практичні потреби вивчення здібностей людей були сформульовані у вигляді важливої проблеми дослідження індивідуальних здібностей та відмінностей. Протягом цього часу накопичено великий досвід використання тестових технологій в освіті. До найбільш розвинених щодо тестування належать такі країни, як: США, Нідерланди, Англія, Японія, Данія. Франція, Ізраїль, Фінляндія. Канада, Австралія, Нова Зеландія.

Розглядаючи історію та тенденції розвитку педагогічних вимірювань як за кордоном, так і в Україні, висвітлимо чинники, які обумовили нинішній стан [10]. Надзвичайно важливим кроком є розпочате у 1993–1994 рр. впровадження тестування на випускних іспитах у середніх школах, результати яких зараховувалися як складова частина вступних іспитів до вищих навчальних закладів України.

Починаючи з 2008 року, усім випускникам для вступу у ВНЗ необхідно складати ЗНО. Це означає, що кожного року аналізують його результати, роблять висновки та прогнози, критикують з метою вдосконалення. Ця спроба стандартизувати в масштабах країни вимірювання й оцінювання рівня знань усіх випускників середніх шкіл за єдиними об’єктивними критеріями є надзвичайно важливою.

1.2 Основні поняття тестології

У цьому та наступному пунктах будемо опиратись на означення із посилання [1,11-14].

Означення 1.1 Тестологія – наука про вимірювання психофізіологічних та особистісних характеристик, а також обсягу та якості знань, умінь, навичок.

Означення 1.2 Тест – сукупність завдань, які переважно вимагають однозначної відповіді, укладений за певними правилами та процедурами, передбачає попередню експериментальну перевірку й відповідає таким характеристикам ефективності, як валідність і надійність.

Означення 1.3 Тестування – метод діагностики із застосуванням стандартизованих запитань та завдань, що мають певну шкалу значень.

До нього вдаються для стандартизованого визначення індивідуальних відмінностей особистості в усьому світі. Вони дають змогу з певною ймовірністю визначити рівень розвитку в індивіда психологічних властивостей (пам’яті, мислення, уяви та ін.), особистісних характеристик, ступінь готовності до певної діяльності, засвоєння знань і навичок.

Ключовим поняттям тестології є поняття «педагогічний тест».

Означення 1.4 Педагогічний тест – а) така система завдань, результат виконання яких групою претендентів дозволяє досить надійно ранжувати їх (надати їм порядкові номери) за якістю навчання, кількістю наявних знань;

б) система стандартизованих завдань, результат виконання яких дозволяє за заданим ступенем точності виміряти знання, навички та вміння випробуваного.

Означення 1.5 Педагогічне тестування – сукупність організаційних і методичних заходів, об’єднаних спільною метою з педагогічним тестом і призначених для підготовки та проведення формалізованих процедур пред’явлення тесту, обробки і представлення результатів його виконання.

Означення 1.6 Банк тестових матеріалів – сукупність систематизованих тестових завдань і педагогічних тестів, що пройшли апробацію і мають відомі характеристики.

До характеристик відносяться як якісні характеристики, що відображають зміст тестового завдання або тесту в цілому (навчальний предмет, розділ, тема, контрольовані вміння і т.д.), так і кількісні (складність тестових завдань, надійність тесту і ін.).

Означення 1.7 Ключ тестового завдання – правильна відповідь на тестове завдання.

Означення 1.8 Дистрактор – варіант відповіді на тестове завдання закритої форми, схожий на правильний, але який не є таким.

1.3 Загальні вимоги до якості тестів

Дидактичні можливості тестового контролю можуть бути реалізовані за умови виконання певних вимог до складання тесту (контрольного завдання, програми) [13].

Означення 1.9 Валідність – це придатність тесту для виміру саме тієї властивості, яку він оцінює.

Валідність тесту дає відповідь на питання про те, що вимірює тест, чи відповідає він меті, для якої застосовується. Валідність тесту означає, що за його допомогою вимірюються саме ті знання, уміння та навички, які він призначений оцінювати. Валідність тесту визначають за трьома характеристиками*:*

а) функціональність – дії студентів у процесі виконання тестів повинні відповідати за більшістю показників тим операціям, які ці контрольні завдання перевіряють;

б) змістовність – для виконання тесту студент застосовує знання того навчального матеріалу, рівень засвоєння якого цей тест перевіряє;

в) прогностичність – інформація, що одержана під час аналізу результатів виконання тесту, повинна містити достовірні показники для визначення змісту та прогнозування результатів наступної роботи.

Означення 1.10 Надійність тесту – це міра стійкості результатів, що впливає на точність виміру конкретної ознаки [11].

Ця характеристика виявляється, насамперед, в одержанні однакових результатів після повторних вимірів. Ступінь надійності залежить від об’єктивності самого тесту, параметрів засобу вимірювання, стабільності ознаки, яку оцінюють. Іншими словами надійність контрольного завдання розуміють як ступінь точності визначення тої чи іншої ознаки, тобто виявлення, наскільки можна довіряти результатам конкретного тесту. Наприклад, тест можна вважати надійним, якщо в усіх випадках перевірки завдання чи його варіантів виявиться, що студенти під час розподілу за показниками успішності займають одні й ті ж місця.

Надійність тесту залежить також від кількості тестових завдань. Вважають, що збільшення кількості контрольних завдань підвищує їхню надійність.

Означення 1.11 Репрезентативність – відповідність характеристик вибірки властивостям генеральної сукупності в цілому [13].

У цьому випадку репрезентативна вибірка — це вибірка тестових питань, в якій їхня кількість представлена в тій же пропорції, що і навчальний матеріал в певній дисципліні. Тест репрезентативний тоді, коли використана під час його розробки вибірка обґрунтована та властиві їй характеристики досить рівномірно розподілені в генеральній сукупності [13].

Занадто мала кількість питань не дасть отримати повну характеристику знань студентів, а неправильний розподіл кількості питань за темами не дасть можливість отримати об’єктивну оцінку знань студентів. Для достовірної перевірки знань студентів необхідно створювати репрезентативні тести, тобто набір питань, що охоплює всю необхідну інформацію із конкретного фрагмента дисципліни: теми, навчального модуля, підсумкового контролю, вступного іспиту тощо.

З репрезентативністю тесту зв’язана його адаптованість, що означає адекватність тесту новим умовам застосування, зокрема іншому соціокультурному середовищі. Адаптація тесту означає аналіз його придатності до застосування в нових умовах, переклад завдання та інструкцій до нього на мову користувача, перевірку його валідності та надійності в нових умовах

застосування, а також випробування (стандартизацію) на великий кількості тестованих.

Проблема адаптованості тестів набула особливої актуальності в останні роки в зв’язку із широким запозиченням західних тестових методик.

Означення 1.12 Дискримінативність – один із критеріїв оцінювання якості тестових завдань, що показує здатність тесту розподіляти учасників відповідно до рівня успішності, виконання діяльності, диференціювати їх відносно максимального і мінімального результатів. Якщо всі досліджувані обирають у тестовому завданні одну і ту саму відповідь, це означає, що завдання не наділене дискримінативністю [12].

Розрізняють змістовну та формальну контрастність. Змістовна контрастність визначається рівнем змістовної відмінності між варіантами відповідей. Ця контрастність різко збільшується за умови застосування поряд із правильними відповідями вочевидь помилкових або таких, що не мають стосунку до завдання. Змістовна контрастність тим менша, чим тонші розбіжності між варіантами відповідей, тобто чим правдоподібніше виглядають дистрактори.

Формальна контрастність визначається рівнем відмінності форми варіантів відповідей. Вона велика, якщо форма правильної відповіді значно відрізняється від форми неправильних відповідей (довжина, повнота), і мала, якщо такі відмінності мінімальні або зовсім відсутні. Для того, щоб тестове завдання виконувало свої функції, дистрактори повинні мати малу змістовну та формальну контрастність.

Означення 1.13 Складність завдання тесту – характеристика завдання тесту, що відображає статистичний рівень спроможності його розв’язання в конкретній вибірці стандартизації [14].

Показником складності тестового завдання є частка вибірки досліджуваних, які не розв’язали завдання. Наприклад, якщо лише 20 % учасників виконали завдання, його можна вважати складним для даної вибірки, якщо 80 % – легким.

1.4 Висновок до першого розділу

Процес перевірки знань та здібностей проводився ще з давніх часів. Використовувались різноманітні випробування та проводились спеціальні іспити для відбору на посади державних служб в стародавній Греції, Спарті та Китаю. Така діагностика стала фундаментом для сучасних тестів.

Вперше термін «педагогічний тест» використав Джордж Фішер перевіряючи знання студентів за допомоги спеціальних книг, які вважаються першими шкільними тестами успішності. Величезний вклад в розвиток теорії тестів зробив англійський психолог Френсіс Гальтон, який визначив три основні її принципи: застосування серії однакових випробувань до великої кількості випробовуваних; статистична обробка результатів; виділення еталонів оцінки.

Першим дослідником та автором поняття «інтелектуальний тест» став американський психолог, один з перших фахівців з експериментальної психології в США, Джеймс Кеттел. На основі отриманих результатів експериментів ним було розроблено набори завдань, які він назвав «розумовими тестами», а також створено систему вимог до процесу тестування.

Розвиток тестології пов’язаний з такими прізвищами, як Дж. Райс, М. Стоун, А. Біне, Т. Симон, В. Макколл, Е. Торндайк, Р. Йеркс, Г. Луфброу, С Пріссей, Ф. Робінсон, Р. Тайлер, Дж. Мак та В. Хант. Усі вони досліджували різні питання теорії тестів, створювали нові типи тестів або удосконалювали існуючі.

Тестологія як наука має основні означення такі, як тест, тестування, педагогічний тест, педагогічне тестування, банк тестових завдань, ключ та дистрактори.

Не кожні тестові завдання можна вважати тестом. Адже тест повинен відповідати таким характеристикам: надійність, ефективність, валідність, репрезентативність, контрастність та дискримінативність.

2 ОСНОВИ ТЕОРІЙ АНАЛІЗУ ЯКОСТІ ТЕСТУВАННЯ ТА АВТОМАТИЗОВАНІ СИСТЕМИ ЗАВНОВАНІ НА ПОЛОЖЕННЯХ ЦИХ ТЕОРІЙ

2.1 Основні положення класичної теорії аналізу якості тестування

Після проведення тестування, традиційно для аналізу якості тесту використовують дві основні теорій обробки результатів тестування – CRT та IRT [13]. Основоположником CRT є відомий британський психолог Чальз Едвард Спірмен. Але всебічно і повно класична теорія тестів вперше викладена в фундаментальній праці американського психолога Гарольда Гулліксена лише 1950 році. З тих пір теорія кілька видозмінювалася, зокрема удосконалювався математичний апарат.

Дослідженням та описом класичної теорії обробки результатів у наш час займаються такі вітчизняні вчені, як Аванесов В. С. [14], Майоров О. М. [15], Челишкова М. Б. [16], Кухар Л. О. [11]

Класична теорія тестів ґрунтується на наступних п’яти основних положеннях [17]:

а) емпірично отриманий результат вимірювання є сумою істинної оцінки та похибки вимірювання:

$$X = T + E, \tag{2.1}$$

де X – тестова оцінка, що спостерігається, тобто отримується емпіричним шляхом, T – істинна оцінка індивідуума, E – оцінка похибки вимірювання.

Цю формулу вважають класичною моделлю істинної оцінки.

б) істинний результат вимірювання (істинну оцінку) можна виразити як математичне сподівання $M(X)$:

$$T = M(X). \tag{2.2}$$

Кожного разу, коли випробуваний виконує тест, його оцінку за цей тест можна вважати значенням випадкової величини. Припустимо, що тест має 50 завдань. Оцінки випробуваних можуть знаходиться в межах від 0 до 50. На цю оцінку можуть впливати систематичні та випадкові помилки. Якщо уявити ідеальну ситуацію, коли випробуваний виконує цей тест скільки завгодно разів (він не втомлюється і повністю забуває завдання після попереднього тестування), то отримувані оцінки з'являтимуться з певною частотою, яка може використовуватись для підрахунку ймовірності появи оцінки за тест. У цьому сенсі тестову оцінку X можна розглянути як значення випадкової величини. Якщо випадкова величина X задана розподілом (табл. 2.1).

Таблиця 2.1 – Розподіл тестової оцінки X

X	x_1	x_2	…..	x_n
P	p_1	p_2	…..	p_n

то її математичне сподівання обчислюється за формулою:

$$M(X) = \sum_{j=1}^{n} x_j * p_j, \tag{2.3}$$

де x_j – j-те значення X (одна із оцінок за тест), p_j – ймовірність отримання оцінки x_j.

Для i-го учасника тестування отримуємо:

$$T_i = M(X_i) = \sum_{j=1}^{n} x_{ij} * p_{ij}, \tag{2.4}$$

Останню формулу можна трактувати наступним чином. Для кожного випробовуваного математичного сподівання можна розглядати як середнє всіх тестових оцінок, які він міг би отримати при багатократному виконанні тесту. Тоді його істинну оцінку можна інтерпретувати як середнє оцінок, отриманих по скінченному числу тестувань при використанні одного і того ж тесту. Істинна

оцінка є статистичним поняттям, що базується на очікуваному значенні, отриманому в даному процесі вимірювання.

в) середнє значення похибок оцінок для генеральної сукупності випробуваних дорівнює нулю:

$$M(E) = 0.$$

г) кореляція між істинною оцінкою та її похибкою для генеральної сукупності випробуваних дорівнює нулю:

$$\rho_{TE} = 0.$$

д) кореляція між похибками оцінок двох будь-яких тестів дорівнює нулю:

$$\rho_{E1E2} = 0.$$

Серед розглянутих п’яти положень перші два є означеннями. Решту три можна отримати логічними міркуваннями із означень. Положення в)-д) ще називають припущеннями або аксіомами моделі. Вони описують ті основні властивості істинних оцінок та їх похибок, які задовольняють застосувати класичну модель істинної оцінки до дослідження надійності тестових балів [3].

Крім цього, основу класичної теорії тестів складають два визначення – паралельні та еквівалентні тести.

Паралельні тести повинні відповідати вимогам (1-5), справжні компоненти одного тесту (T_1) повинні бути рівні істинним компонентів іншого тесту (T_2) в кожній вибірці випробовуваних, що відповідають на обидва тести. Передбачається, що $T_1 = T_2$ і, крім того, рівні дисперсії $S_1^2 = S_2^2$.

Еквівалентні тести повинні відповідати всім вимогам паралельних тестів за винятком одного: справжні компоненти одного тесту не обов’язково повинні рівнятися істинним компонентів іншого паралельного тесту, але відрізнятися вони повинні на одну і ту ж константу c.

Умова еквівалентності двох тестів записується формулою:

$$T_1 = T_2 + c_{12},$$

де c_{12} – константа відмінностей результатів першого і другого тестів.

На основі наведених положень побудована теорія надійності тестів [7].

Приймемо в якості вихідного положення наступне твердження:

$$S_X^2 = S_T^2 + S_E^2,$$

тобто, дисперсія отриманих тестових балів дорівнює сумі дисперсій істинних і помилкових компонентів.

Перепишемо цей вираз в наступному вигляді:

$$\frac{S_T^2}{S_X^2} = 1 - \frac{S_E^2}{S_X^2}.$$

Права частина цієї рівності є надійність тесту, позначимо через r. Таким чином надійність тесту можна записати у вигляді:

$$r = 1 - \frac{S_E^2}{S_X^2}.$$

На основі цієї формули в подальшому були запропоновані різні вирази для знаходження коефіцієнта надійності тесту. Надійність тесту є найважливішою характеристикою його якості. Якщо невідома надійність, то результати тестування неможливо інтерпретувати. Надійність тесту характеризує його точність як вимірювального інструмента. Висока надійність означає високу повторюваність результатів тестування в однакових умовах.

У класичній теорії тестів найважливішою проблемою є визначення істинної оцінки випробуваного (T). Емпіричний тестовий бал (X) залежить від багатьох умов – рівня складності завдань, рівня підготовленості випробуваних,

кількості завдань, умов проведення тестування і т.д. У групі сильних, добре підготовлених випробуваних, результати тестування будуть як правило, краще, ніж в групі слабо підготовлених випробуваних. У зв’язку з цим залишається відкритим питання про величину міри складності завдань на генеральної сукупності досліджуваних. Проблема полягає в тому, що реальні емпіричні дані отримують на зовсім не випадкових вибірках випробовуваних. Як правило, це навчальні групи, що представляють собою безліч учнів досить сильно взаємодіючих між собою в процесі навчання і навчаються в умовах, часто не повторюються для інших груп.

Відповідно до офіційних звітів Українського центру оцінювання якості освіти про проведення зовнішнього незалежного оцінювання результатів навчання оцінювання якості тестів ЗНО відбувається із застосування CRT [14].

У звіті обчислюються основні характеристики вибірки балів учасників тестування: середнє, мода, медіана, стандартне відхилення, асиметрія та ексцес. Оцінюються такі характеристики тесту як складність тесту, розподільна здатність тесту, надійність тесту за коефіцієнтом Кронбаха та стандартна похибка вимірювань. Також будується розподіл учасників тестування за кількістю набраних балів. Оцінюються такі характеристики окремих тестових завдань як складність, дискримінативність (D-index), кореляція між результатами виконання певного завдання тесту учасниками тестування та загальним балом. Надаються рекомендації щодо інтерпретації різних значень цих характеристик. Будуються діаграми розподілів тестових завдань за складністю та розподільною здатністю, точкові діаграми розсіювання тестових завдань за складністю та розподільною здатністю, діаграми розсіювання тестових завдань за коефіцієнтом кореляції.

2.2 Автоматизовані системи обробки результатів на основі CRT

На сьогоднішній день, створена досить велика кількість програм, які допомагають не тільки розробляти та проводити тестування, а й також аналізувати отримані результати на основі двох головних підходів до обробки результатів тестування Item Response Theory та Classical Response Theory. Автоматизовані системи дають змогу опрацьовувати великі об’ємі даних та обчислювати основні статистичні показники: характеристику тесту і тестових завдань, рівень підготовки випробовуваних та ступінь відповідності рівня навчальних досягнень рівню складності завдань.

Проте лише частина розроблених програмних засобів для конструювання тестів безкоштовна, більшість з програм закритого типу для особистого користування.

Одна з лідерів в комп’ютерному тестуванні є Assessment Systems Corporation з її програмами для адаптивного тестування. ASC спеціалізується у галузі комп’ютерного тестування та психометричного програмного забезпечення на основі теорії IRT та CTT.

ASC розробляє програмні забезпечення, які дозволяє публікувати звіти про професійний психометричний аналіз, використовуючи Classical Test Theory. До найбільш відомих програм Assessment Systems Corporation в основі яких лежить CTT відносяться Iteman 4, Lertap, CITAS. Програми MyTestXPro, R, SPSS Statistics також використовують класичну теорію обробки результатів тестування [21].

Означення 2.1 MyTestXPro – це система програм для створення і проведення комп’ютерного тестування знань, виставлення оцінки за вказаною в тесті шкалою, збору і аналізу результатів [18].

За допомогою програми MyTestXPro можлива організація і проведення тестування, іспитів в будь-яких освітніх установах (вузи, коледжі, школи) як з метою виявити рівень знань по будь-яким навчальним дисциплінам, так і з навчальними цілями. Підприємства та організації можуть здійснювати атестацію та сертифікацію своїх співробітників.

Програма складається з трьох модулів: Модуль тестування, Редактор тестів і Журнал тестування. Використовуючи Журнал тестування MyTestXPro, можна отримати детальні результати кожного тесту.

Вибрати які саме стовпці будуть відображені в таблиці результатів, для цього потрібно обрати команду Дії → Результати → Показувати колонки ... (рис. 2.1). Ця установка може бути збережена для подальшого використання, тобто при наступному запуску будуть показані обрані вами колонки.

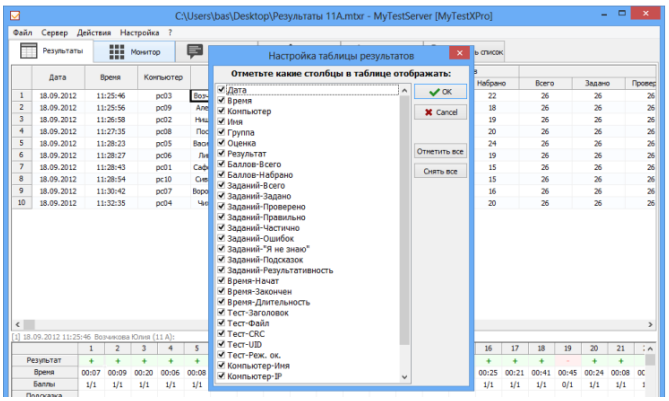


Рисунок 2.1 – Вікно налаштувань таблиці отриманих результатів

Зібрані результати можна проаналізувати спільно. Для цього обираємо команду Дії → Результати → Аналіз → Вибрати з усіх ... (або Вибрати із виділених ...) як показано на рис. 2.2.

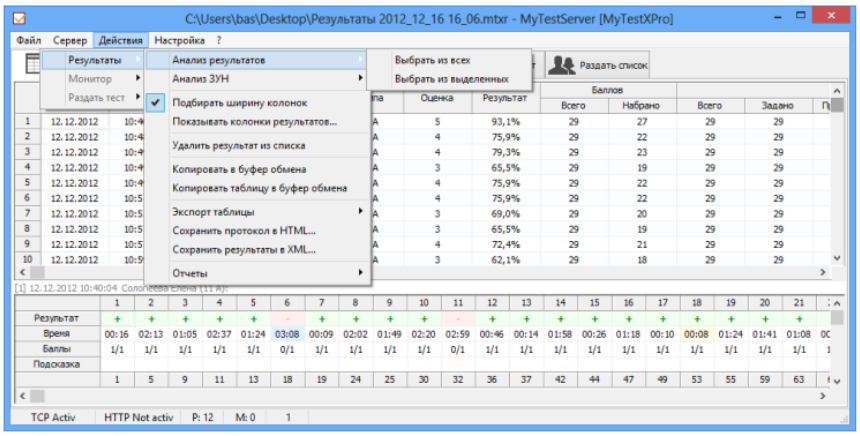
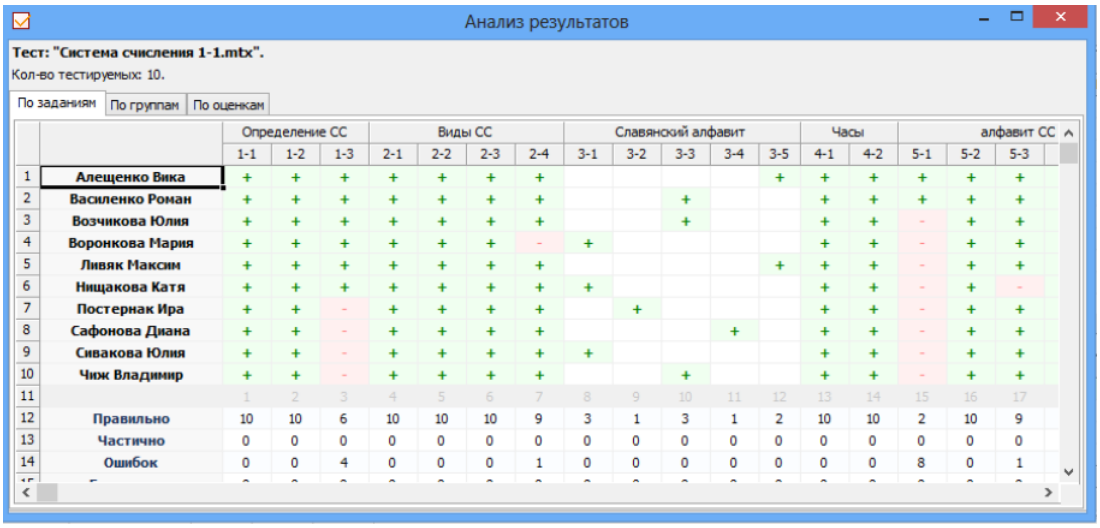


Рисунок 2.2 – Вікно для аналізу результатів

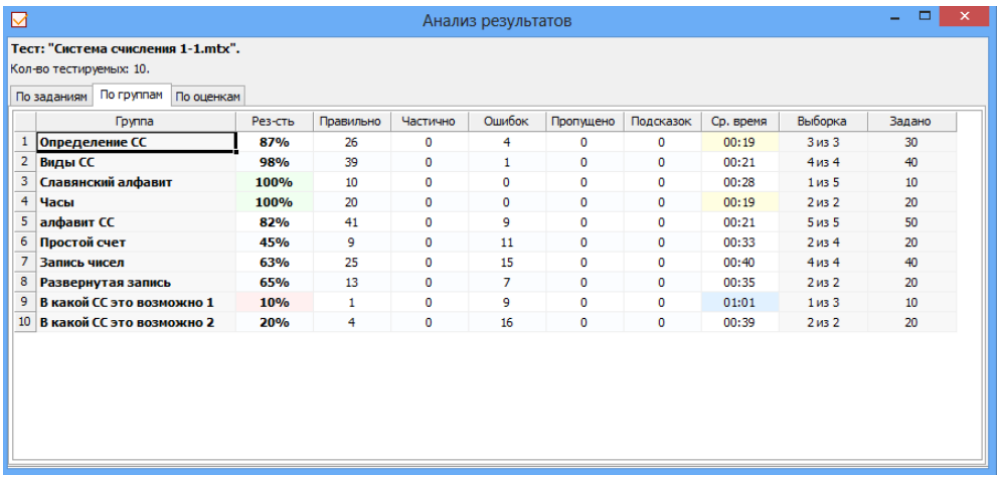
Аналіз тестування за завданнями дозволяє отримати таблицю, стовпці якої відповідають номерам завдань у тесті, а верхні рядки таблиці – кожному учневі, що проходив цей тест, нижні рядки загальну статистичну інформацію за завданнями. Комірки таблиці виділяються різними кольорами залежно від їх значень (рис. 2.3). Це дозволяє більш швидко проаналізувати результати.



		Определение СС			Виды СС				Славянский алфавит					Часы		алфавит СС		
		1-1	1-2	1-3	2-1	2-2	2-3	2-4	3-1	3-2	3-3	3-4	3-5	4-1	4-2	5-1	5-2	5-3
1	Алещенко Вика	+	+	+	+	+	+	+					+	+	+	+	+	+
2	Василенко Роман	+	+	+	+	+	+	+			+			+	+	+	+	+
3	Возчикова Юлия	+	+	+	+	+	+	+			+			+	+	-	+	+
4	Воронкова Мария	+	+	+	+	+	+	-	+					+	+	-	+	+
5	Ливяк Максим	+	+	+	+	+	+	+					+	+	+	-	+	+
6	Нищакова Катя	+	+	+	+	+	+	+	+					+	+	-	+	-
7	Постернак Ира	+	+	-	+	+	+	+		+				+	+	-	+	+
8	Сафонова Диана	+	+	-	+	+	+	+				+		+	+	-	+	+
9	Сивакова Юлия	+	+	-	+	+	+	+	+					+	+	-	+	+
10	Чиж Владимир	+	+	-	+	+	+	+			+			+	+	-	+	+
11		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
12	Правильно	10	10	6	10	10	10	9	3	1	3	1	2	10	10	2	10	9
13	Частично	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	Ошибок	0	0	4	0	0	0	1	0	0	0	0	0	0	0	8	0	1

Рисунок 2.3 – Аналіз результатів кожного випробуваного

Аналіз тестування за групами дозволяє дізнатися результативність по кожній групі завдань тесту. Кожен рядок таблиці відповідає окремій групі в тесті (рис. 2.4). Таким чином можна з’ясувати, наприклад, завдання яких груп викликають найбільші труднощі в учнів і скоригувати процес навчання.



	Группа	Рез-сть	Правильно	Частично	Ошибок	Пропущено	Подсказок	Ср. время	Выборка	Задано
1	Определение СС	87%	26	0	4	0	0	00:19	3 из 3	30
2	Виды СС	98%	39	0	1	0	0	00:21	4 из 4	40
3	Славянский алфавит	100%	10	0	0	0	0	00:28	1 из 5	10
4	Часы	100%	20	0	0	0	0	00:19	2 из 2	20
5	алфавит СС	82%	41	0	9	0	0	00:21	5 из 5	50
6	Простой счет	45%	9	0	11	0	0	00:33	2 из 4	20
7	Запись чисел	63%	25	0	15	0	0	00:40	4 из 4	40
8	Развернутая запись	65%	13	0	7	0	0	00:35	2 из 2	20
9	В какой СС это возможно 1	10%	1	0	9	0	0	01:01	1 из 3	10
10	В какой СС это возможно 2	20%	4	0	16	0	0	00:39	2 из 2	20

Рисунок 2.4 – Аналіз тестування за групами завдань

Аналіз за оцінками показує Середній бал, Успішність, Якість знань ... Ці параметри часто потрібно обчислювати після проведення, наприклад, будь-якого «зрізу знань». Програма вирахує їх автоматично (рис. 2.5). Коефіцієнти для обчислення можна задати або змінити в налаштуваннях програми.

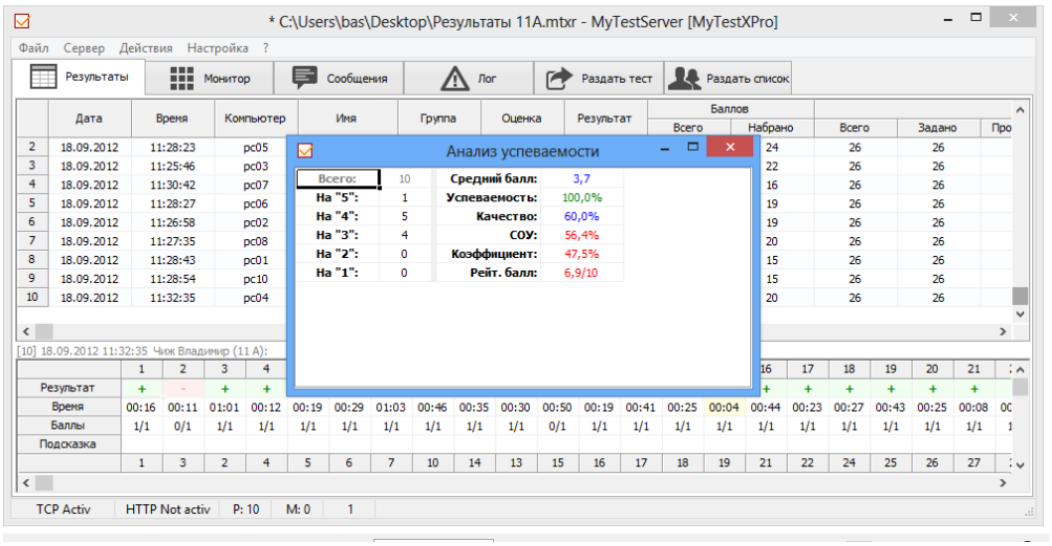


Рисунок 2.5 – Аналіз успішності тесту

Проаналізувавши модуль програми MyTestXPro – Журнал тестування, можна зробити висновки, що за його допомогою визначаються лише базові(примітивні) характеристики якості тесту класичної теорії обробки результатів тестування. Якщо потрібен більше детальний аналіз, отримані дані можна експортувати в xml-файл і досліджувати іншими програмами, або в електронні таблиці Excel для побудови різних діаграм і графіків. Таким чином, програма MyTestXPro дозволяє оцінити успішність тестованого, але не якість тесту.

Означення 2.2 Iteman – додаток Windows, призначений для надання детальних звітів про аналіз предметів та тестів з використанням класичної теорії випробувань (CRT) [19].

Мета цих звітів – допомогти оцінити та покращити якість тестових завдань шляхом вивчення їх психометричних характеристик.

Для дослідження обрано безкоштовну версію Iteman 4, яка обмежена 100 предметами та 100 респондентами. Незважаючи на такі обмеження, цей додаток

досить вдало підходить для викладання або вивчення класичної теорії тестів, а також для реального застосування у багатьох тестах меншого масштабу.

У Iteman є графічний інтерфейс користувача GUI, який дозволяє легко запускати програму, навіть якщо ви не знайомі з психометрикою. Інтерфейс Iteman розділений на чотири вкладки: файли, формат вводу, параметри балів, параметри виводу.

На вкладці «Файли» вказуються файли, які слід використовувати: матриця даних, файл контроль завдань, вихідний файл та необов'язковий зовнішній файл оцінки (рис. 2.6).

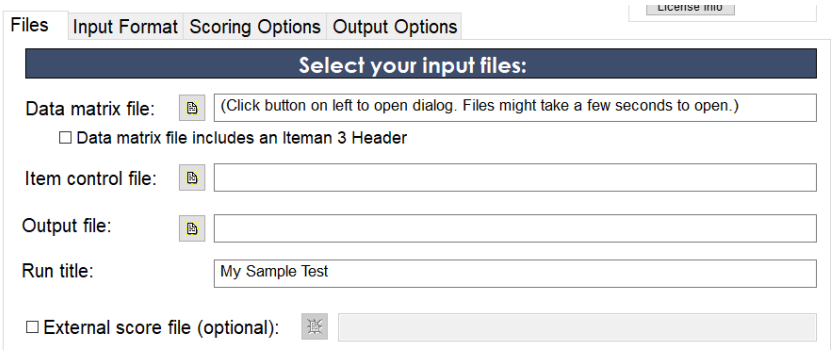


Рисунок 2.6 – Вкладка «Файли» у додатку Iteman

Iteman 4 потрібні два вхідні файли – файл матриці даних та файл контроль завдань. Ці файли можна створити в таких програмах, як Microsoft Excel, Microsoft Word, Текстовий документ. Файл матриці даних – це файл, який містить ідентифікатор або ім'я випробуваного та відповіді на кожне завдання. Відповіді можуть бути подані за допомогою букв (A, B, C, D… або a, b, c, d…) або чисел (1,2,3,4…). Програма дотримується стандартного підходу: рядки – кількість людей, а стовпці – кількість предметів чи спостережень. У програмі існує декілька типів файлів матриць даних – розмежований та фіксована ширина. При першому типі Iteman 4 дозволяє використовувати файл матриці отриманих даних, який обмежений комою або табуляцією (рис. 2.7). Значення, розділені комами простий та зручний у використані, оскільки надає можливість редагувати файли в стандартному програмному забезпеченні.

```
Person9,4,2,1,3,3,2,3,4,1,2
Person10,1,2,1,3,3,2,3,4,1,0
Person11,3,3,2,3,1,2,3,4,1,3
Person12,1,2,2,3,3,2,3,4,1,4
Person13,2,2,1,4,3,2,3,4,1,1
```

Рисунок 2.7 – Матриця вхідних даних з розмежуванням

Відповіді на кожне завдання розмежовані комами, а це означає, що його можна створити в програмі електронних таблиць, а потім «Зберегти як» текстовий файл з обмеженими вкладками або CSV.

Якщо було запропоновано аналіз функціонування диференціальних елементів (DIF), код членства в групі DIF повинен відповідати ідентифікатору випробуваного, як показано на рисунку 2.8. Важливо зазначити, що коди членства DIF (M і F) не будуть розпізнані, якщо вони будуть включені до складу ідентифікатора, наприклад, Person9M.

```
Person9,M,4,2,1,3,3,2,3,4,1,2
Person10,F,1,2,1,3,3,2,3,4,1,0
Person11,M,3,3,2,3,1,2,3,4,1,3
Person12,F,1,2,2,3,3,2,3,4,1,4
Person13,F,2,2,1,4,3,2,3,4,1,1
```

Рисунок 2.8 – Матриця вхідних даних з диференціальними елементами

Підхід із фіксованою шириною – це текстовий файл, у якому всі стовпці повинні бути точно вирівняні. Приклад цього показано на рисунку 2.9 для 10 предметів та 5 досліджуваних. У цьому файлі 9 стовпців ідентифікатора (останні два порожні) та 10 стовпців відповідей. Додаткові стовпці можна ігнорувати, тому видаляти дані не потрібно.

```
Person1   4213323412
Person2   1213323410
Person3   3323123413
Person4   1223323414
Person5   2214323411
```

Рисунок 2.9 – Матриця вхідних даних з фіксованою шириною

Файл контроль завдань – файл який містить ідентифікатор номера завдань, правильний варіант відповіді, кількість варіантів відповідей, домен або область вмісту, статус включення (Y = так, N = ні, P = предтест), тип завдання. Поняття «домен» може включати зміст програми з предмета, вимір рівнів засвоєння, вимоги освітнього стандарту, компетентність тощо. У предметному тесті визначення домену – це відповідність або, власне, кінцевий результат вивчення теми, розділу, предмета, відповідність засвоєння етапу навчальної програми та висновки на основі одержаних результатів.

Приклад файлу контроль показаний на рисунку 2.10. Є десять завдань, дев'ять з яких мають декілька варіантів вибору та одна часткова позичка. Перші п'ять розміщені у домені 1, а останні п'ять – у домені 2. Перші чотири завдання у кожному домені зараховуються, тоді як п'ятий елемент у кожному –має статус попереднього тесту. Кожне завдання має 4 варіанта відповіді. Відповіді можуть бути буквені або числовими. Завдання 7 має дві правильні відповіді 3 та 1. Для нього відповіді будуть оцінені як правильні, якщо випробуваний відповість або 3, або 1.

Item01	1	4	Science	Y	M
Item02	2	4	Science	Y	M
Item03	3	4	Science	Y	M
Item04	4	4	Science	Y	M
Item05	1	4	Science	P	M
Item06	2	4	Reading	Y	M
Item07	31	4	Reading	Y	M
Item08	4	4	Reading	Y	M
Item09	1	4	Reading	Y	M
Item10	+	5	Reading	P	P

Рисунок 2.10 – Файл контроль завдань

У контрольному файлі повинно бути стільки рядків, скільки завдань у тесті. Програма підраховує рядки інформації у контрольному файлі, що слугує загальною кількістю елементів у тесті.

Після створення файлів матриці даних та контролю завдань, залишається створити пустий вихідний файл та розмістити його у третьому полі. Вихідний файл повинен мати розширення .docx.

У вкладці «Формат вводу» обирається тип форматування даних (рис. 2.11). Для фіксованої ширини визначається кількість стовпців файла матриці даних для ідентифікаторів та відповідей завдань і дозволяє вказати характер коду, який використовується в матриці даних для вказівки пропущених завдань. Якщо ж тип файлу матриці розмежований, то потрібно обрати вид розмежування – кома або табуляція.

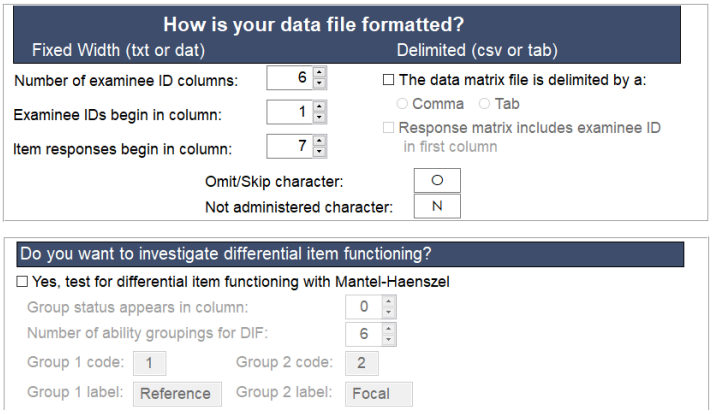


Рисунок 2.11 – Вкладка «Формат вводу»

Вкладка «Параметри балів» дає змогу виконувати масштабний підрахунок та проводити дихотомічну класифікацію.

Якщо ваша програма тестування звітує про масштабні бали на основі балів, що не відповідають кількості правильних відповідей, їх можна розрахувати безпосередньо. Масштабовані бали обчислюються за допомогою функції масштабування для загальної кількості правильних балів та/або балів, що відповідають правильному номеру домену.

Масштабований бал часто використовується для маскування деталей про тест, таких як точна кількість предметів або для вираження балів за іншою шкалою, ніж кількість правильних. У більшості масштабних тестів використовується форма масштабування балів. Існує кілька видів масштабування (рис. 2.12):

а) лінійне масштабування: необроблені бали спочатку множать на коефіцієнт нахилу, потім додають мінімальне значення нової шкали;

Наприклад, якщо ви хочете, щоб результати оцінювались за шкалою від 100 до 200 для тестування 50 предметів, масштабний бал може бути визначений як $SCALE = RAW * 2 + 100$.

б) стандартизоване масштабування: необроблені бали перетворюються на середнє значення X і стандартне відхилення Y.

Ця форма масштабування доречна, якщо ви хочете відцентрувати середнє значення тесту навколо постійного значення для використання у звіті. Наприклад, класична шкала IQ із середнім значенням 100 та стандартним відхиленням 15.

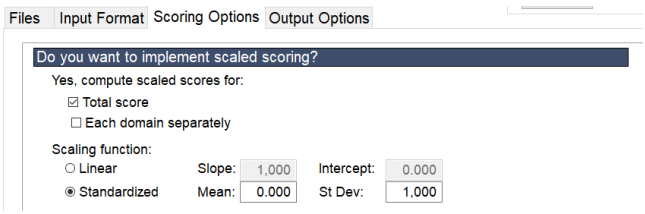


Рисунок 2.12 – Вкладка «Параметри балів»

Вкладка «Параметри виводу» надає можливість налаштувати вихідний звіт під ваші конкретні потреби (рис. 2.13).

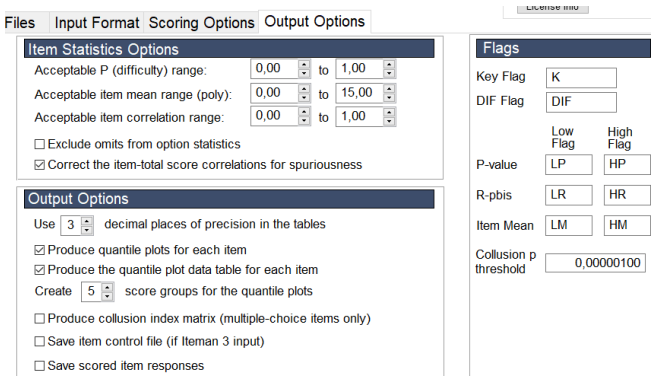


Рисунок 2.13 – Вкладка «Параметри виводу»

У цій вкладці можна обирати вид кореляційних зв’язків, поділ опитуваних на слабку та сильну групи, аналіз тесту у підгрупах, уточнення параметрів завдань для коротких тестів.

Після виконання всіх вимог та налаштувань, проводиться аналіз отриманих даних, натиснувши кнопку «Run». Iteman 4 надає три вихідні файли за замовчуванням: звіт DOCX, файл CSV статистики завдань та файл CSV статистики балів випробуваних. Файл CSV включає ті ж статистичні дані, що і у звіті DOCX, але у формі CSV, щоб ви могли маніпулювати даними в електронній таблиці або легко завантажувати їх у програмне забезпечення, наприклад FastTest.

Первинний результат, звіт DOCX, подається як формальний звіт, який може бути наданий тестуючим розробникам або експертам з предметних питань. Він починається з титульної сторінки, за якою супроводжується зведеною інформацією вхідних специфікацій (рис. 2.14).

Рисунок 2.14 – Титульна сторінка звіту

Далі, у звіті надаються зведені статистичні дані про рівень тестування на основі отриманих балів або зовнішніх балів, якщо вони використовуються. Це робиться для загальної оцінки (усіх завдань), а також для фактичної оцінки (лише відібраних завдань), лише для попереднього тестування та всіх доменів чи областей вмісту. Далі наведено визначення стовпців у таблиці 2.2.

Таблиця 2.2 – Переклад позначень у програмі Iteman 4

Позначення	Пояснення
Items	Кількість завдань у тесті
Mean	Середнє значення
SD	Стандартне відхилення
Min score	Мінімальний бал
Max score	Максимальний бал
Mean P	Середнє значення трудності завдань у тесті
Mean Rpbis	Середнє значення коефіцієнта кореляції точково-бісеріального завдань у тесті

Зведена таблиця 2.3 рівня тесту дозволяє зробити важливі порівняння між різними частинами тесту. Наприклад, чи можна порівняти предтестові завдання за складністю з поточними завданнями? Чи є завдання в Домені 2 складнішими, ніж в Домені 1?

Таблиця 2.3 – Підсумкова таблиця рівня тесту

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	42	38.560	5.288	27	46	0.863	2.020
Scored Items	36	33.600	4.703	23	40	0.869	2.020
Pretest items	6	4.960	1.087	2	6	0.827	0.000
Domain 1	8	7.360	0.776	5	8	0.920	0.000
Domain 2	16	13.600	2.185	7	16	0.850	0.000
Domain 3	12	12.640	2.926	7	17	0.860	2.020

Аналіз надійності надає таблицю, яка підсумовує статистику надійності, обчислену Iteman 4. Коефіцієнт α та стандартна похибка вимірювання SEM (на основі α) обчислюються для всіх завдань, обчислюваних завдань, завдань предтесту, і для кожного домену окремо. Обчислюються три форми надійності з розділеною половиною (табл. 2.4). Спочатку тест випадковим чином ділиться на дві половини, співвідношення Пірсона обчислюється між загальним балом та двома половинами. Також передбачено співвідношення розділеної половини між загальними балами першої половини та другої половини тесту та непарними і парними завданнями в тесті. Оскільки ці кореляції обчислюються, використовуючи половину загальної кількості завдань, надаються також кореговані кореляції Спірмена-Брауна.

Таблиця 2.4 – Аналіз надійності завдань

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
All items	0.765	2.561	0.537	0.473	0.707	0.699	0.643	0.829
Scored items	0.731	2.439	0.462	0.434	0.682	0.632	0.605	0.811
Pretest items	0.519	0.754	-	-	-	-		
Domain 1	0.073	0.747	0.014	0.182	-0.008	0.028	0.308	-0.016
Domain 2	0.642	1.307	0.607	0.380	0.328	0.755	0.551	0.494
Domain 3	0.590	1.874	0.209	0.149	0.600	0.345	0.259	0.750

Після статистичної таблиці тестових рівнів подається гістограма частот індивідуальних балів (рис. 2.15).

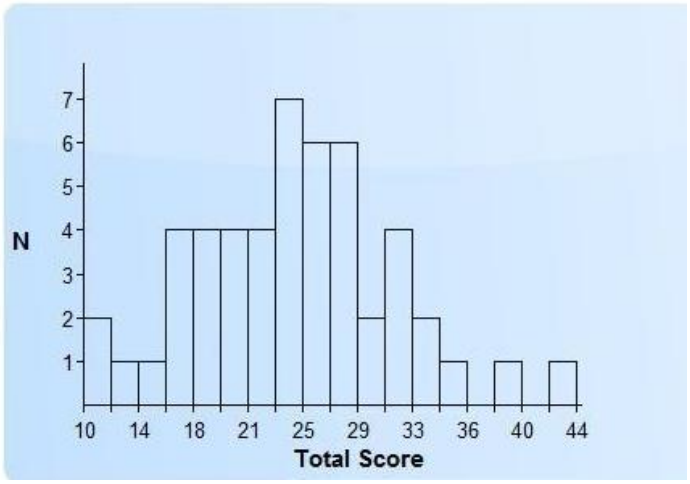


Рисунок 2.15 – Гістограма частот індивідуальних балів

Після гістограми ці дані подані у таблиці. Подібні графіки створюються для кожного домену, якщо є більше одного домену. Також можуть бути подані гістограми складності завдань у тесту, міри дискримінації завдань.

Класична функція CSEM (умовна стандартна помилка вимірювання) це міра помилки прогнозованого балу при повторному проходженні тесту випробуваними. Сюжет обчислюється за допомогою лордової Формули IV. Формула IV CSEM чітко припускає, що всі завдання приймають значення 0 або 1. Зразок графіку CSEM показаний на рис 2.16. Низьке значення означає, що ми очікуємо, що обстежуваний отримає аналогічну оцінку при повторному переході.

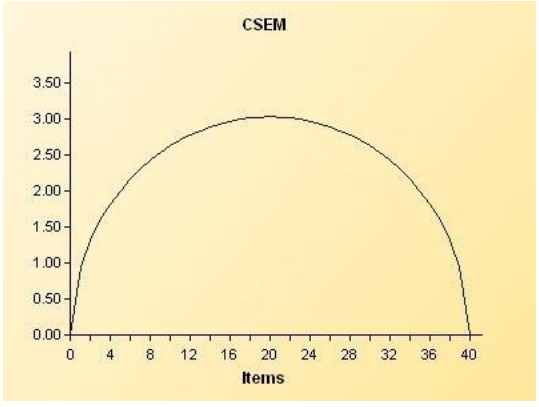


Рисунок 2.15 – Графік функції CSEM

Після статистики рівня тесту надається детальна таблиця статистики для кожного завдання.

Для кожного завдання представлено чотири таблиці:

а) таблиця інформації про завдання – записує інформацію, надану файлом контролю для цього завдання (рис. 2.16);

Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
1	Item01	A	Yes	4	1	LR

Рисунок 2.16 – Таблиця інформації про завдання

б) таблиця статистики завдання – загальна статистика завдань;

Таблиця статистики завдань представляє загальну статистику. Дві найважливіші статистичні дані на рівні дослідження завдання для дихотомічно набраних даних – це значення P та точково-бісеріальна кореляція, які представляють складність та дискримінацію завдання відповідно.

Item statistics

N	P	Domain Rpbis	Domain Rbis	Total Rpbis	Total Rbis	Alpha w/o
60	0,917	0,180	0,324	-0,064	-0,116	0,705

Рисунок 2.17 – Таблиця статистики завдання

Значення P – це частка досліджуваних, які відповіли правильно на завдання. Значення P коливається від 0 до 1. Високе значення (0,95) означає, що завдання легке, а низьке значення (0,25) означає –завдання важке. Точково-бісеріальна кореляція (Rpbis) – це міра дискримінаційної або диференційованої сили завдання. Rpbis коливається від -1 до 1. Негативний Rpbis вказує на погане завдання.

в) статистика варіантів відповідей – детальна статистика для кожного дистрактора, яка допомагає діагностувати проблеми в завданні (рис. 2.18);

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	55	0,917	-0,064	-0,116	29,200	3,584	Maroon	**KEY**
B	4	0,067	0,077	0,148	29,250	4,992	Green	
C	0	0.000	--	--	--	--	Blue	
D	1	0,017	-0,010	-0,031	28,000	0,000	Olive	
Omit	0							
Not Admin	0							

Рисунок 2.18 – Статистична обробка дистракторів у завданні

Таблиця статистики варіантів відповіді представляє статистику для кожного окремого дистрактора. Головне, що слід вивчити в цій частині таблиці, це те, що жодна неправильна відповідь не повинна приймати значення Rpbis більше, ніж правильна. Це буде вказувати на те, що ті, хто написав тест на високий рівень обирають неправильну відповідь.

г) дані квантильного графіку – значення, використані для створення квантильного графіку (рис. 2.19).

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
A	55	0,909	0,923	0,900	1,000	0,895	Maroon	**KEY**
B	4	0,091	0,000	0,100	0,000	0,105	Green	
C	0	0,000	0,000	0,000	0,000	0,000	Blue	
D	1	0,000	0,077	0,000	0,000	0,000	Olive	

Рисунок 2.19 – Дані квантильного графіку

У таблиці даних представлені значення, обчислені для створення квантильного графіку (рис. 2.20). Оскільки він містить ту саму інформацію, сам квантильний сюжет представляє корисну картину ефективності завдання, але цю таблицю можна використовувати для детального вивчення його ефективності, щоб допомогти діагностувати можливі проблеми.

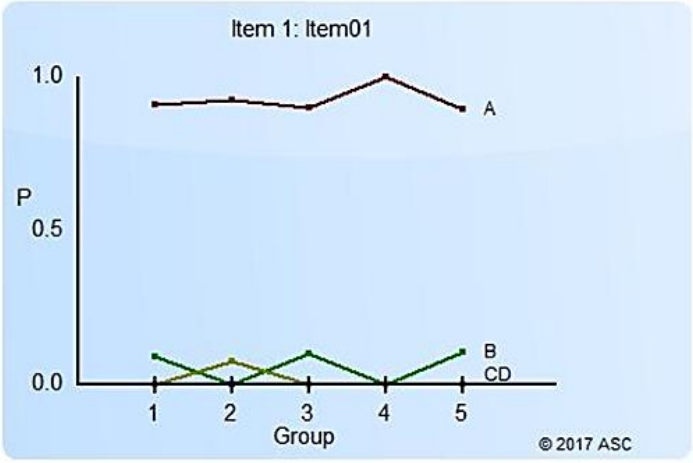


Рисунок 2.20 – Квантильний графік відповідей завдання

Як правило, хороше завдання має позитивний нахил для графіку, що характеризує правильну відповідь, тоді як нахил для неправильних параметрів має бути негативним.

Якщо ви хочете виконати аналіз декількох елементів за один запуск програми, вам слід створити файл декількох запусків (MRF). Наприклад, якщо ви працюєте зі шкільними оцінками і наприкінці року вам пропонують 80 різних тестів для аналізу, але всі вони форматовані однаково, ви можете запустити Iteman один раз, а не 80 разів.

Проаналізувавши основні можливості додатку Iteman 4, можна відмітити, що дана програма оцінює загальні статистичні показники тесту, кожного завдання та кожної альтернативи залежно від обраної опції аналізу. Аналіз характеристик кожного завдання дозволяє виявити ті з них, які не коректно працюють у тесті і потребують вдосконалення.

Дані матриці результатів та файлу контролю завдань потрібно вводити власноруч або ж можна скористатись додатковими програми, але цей варіант стосуються лише матриці даних. На жаль, файл контроль вводиться лише власноруч. Хоча інтерфейс у програми не складний, але для створення файлу контроль необхідно дотримуватись певних вимог, що може викликати незручності у користувача. Таким чином, даний додаток має свої мінуси лише при оформлені матриці даних та файлу контролю, але на противагу цьому, отримуємо детальний звіт щодо якості тесту.

Програма Lertap [20] надає три звіти результатів тестів: повний звіт про статистику (Stats1f), короткий звіт про статистику (Stats1b), і верхній-нижній звіт (Stats1ul). У цих звітах подано рівень знань випробуваних, дискримінації предметів та надійності тесту. Також Lertap надає показники, необхідні для формування інтервалів довіри, діапазони балів, які відображають неточність тестів, що дає можливість для покращення тесту в подальшому використанні.

2.3 Основні положення IRT теорії аналізу якості тестування

На сьогоднішній день особливої актуальності набуває аналіз якості тесту за допомогою сучасної теорії IRT, що спрямована на оцінювання латентних якостей випробуваного та параметрів завдань тесту на основі математико-статистичних моделей вимірювання [21].

Найбільш значних успіхів у розвитку сучасної теорії домоглися Фредерік Лорд, що вважається її засновником, Георг Раш, автор так званої «Rasch measurement», яку називають однопараметричною моделлю IRT, Аллан Бірнбаум, автор дво- і трипараметричної моделей. У ході подальшого розвитку IRT В. Аванесовим була запропонована чотирьохпараметрична модель [22].

В основу даної теорії покладено припущення про існування взаємозв'язку між спостережуваними результатами і латентними якостями тих, хто виконує тест.

Означення 2.3 Латентний параметр – це параметр, що показує здібності особистості, які недоступні для прямого спостереження [21].

У сучасній теорії встановлюється зв'язок між двома множинами значеннями латентних параметрів. Перша множина визначає рівень підготовки випробовуваних θ_i $(i = 1,2, \dots, N)$, а друга множина описує складність j-го завдання β_j $(j = 1, 2, \dots, n)$. Георг Раш створив математичну модель зв'язку між латентними параметрами при умові, що параметри оцінюються в одній шкалі, а це в свою чергу дозволяє вимірювати рівень досягнень учасника в спеціальних одиницях виміру, які називаються логітах.

Використовуючи IRT, розглядають умовну ймовірність P_i вірного виконання i-тим випробовуваним з рівнем підготовки θ_i, різних за складність завдань тесту, при цьому вважають θ_i параметром, а β – незалежною змінною. Аналогічно вводиться P_j для визначення ймовірності правильного виконання j-го завдання складності β_j різними випробуваними групи, де θ – незалежна змінна, а β_j – параметр, що визначає складність j-го завдання тесту. Графік функції P_i має назву індивідуальної кривої i-го випробуваного, а графік функції P_j – характеристичної кривої j-го завдання.

У IRT теорії основними моделями є однопараметрична модель Г. Раша, двопараметрична та трипараметрична моделі А. Бірнбаума. Однопараметрична модель Г. Раша, яку часто називають логістичною моделлю, описана формулою:

$$P_j(\theta) = \frac{e^{1,7(\theta-\beta_j)}}{1+e^{1,7(\theta-\beta_j)}}. \qquad (2.5)$$

Застосовуючи однопараметричну модель Г.Раша, встановлюємо ймовірність виконання завдання, знаючи рівень підготовленості учасника тестування та рівень складності цього завдання. Але в цій моделі коефіцієнт крутизни кривих завдань є однаковим значенням, через що можна невірно інтерпретувати отримані результати.

Щоб уникнути даної проблеми використовують двопараметричну модель А. Бірнбаума, яка включає в себе параметр диференційованої спроможності завдання тесту, що дозволяє розрізняти учасників тестування з різним рівнем навчальних досягнень. Двопараметрична модель А.Бірнбаума для умовної ймовірності вірного виконання завдання тесту випробовуваними знаходиться за формулою:

$$P_j(\theta) = \frac{e^{1,7a_j(\theta-\beta_j)}}{1+e^{1,7a_j(\theta-\beta_j)}}, \qquad (2.6)$$

де a_j – параметр характеристики диференційованої спроможності завдання при зміні різних значень θ. Його можна отримати за формулою:

$$a_j = \frac{(r_{bis})_j}{\sqrt{1-(r_{bis})_j^2}}, \tag{2.7}$$

де $(r_{bis})_j$ – бісеріальний коефіцієнт кореляції, який знаходиться за формулою:

$$(r_{bis})_j = \frac{(\bar{X}_1)_j - (\bar{X}_0)_j}{S_x} \cdot \frac{(N_1)_j - (N_0)_j}{uN\sqrt{N^2-N}}, \tag{2.8}$$

де $(\bar{X}_1)_j$ – середнє значення індивідуальних балів випробовуваних, які правильно виконали j-те завдання тесту, $(\bar{X}_0)_j$ – середнє значення індивідуальних балів випробовуваних, які виконали невірно j-те завдання тесту, S_x – стандартне відхилення за множиною значень індивідуальних балів, $(N_1)_j$ – число випробовуваних, які правильно виконали j-те завдання, $(N_0)_j$ – число випробовуваних, які виконали невірно j-те завдання, N – загальна кількість випробовуваних $(N = N_1 + N_0)$, u – ордината нормованого нормального розподілу в точці, за якої лежить $100 \cdot \frac{N_1}{N}\%$ площі під нормальною кривою.

Також існує трипараметрична модель А. Бірнбаума, яка враховує ймовірність вірної відповіді на завдання в тому випадку, якщо відповідь була вгадана, а не заснована на знаннях випробовуваного. Трипараметрична логістична модель А. Бірнбаума ймовірності вірної відповіді випробовуваним на j-е завдання тесту знаходиться за формулою:

$$P_j\{x_{ij} = 1 \mid \beta_j\} = c_j + (1-c_j)\frac{e^{1{,}7a_j(\theta-\beta_j)}}{1+e^{1{,}7a_j(\theta-\beta_j)}}, \tag{2.9}$$

де c_j – параметр вгадування.

Параметр c_j визначається кількістю відповідей до закритих завдань тесту. Для завдання з п'ятьма відповідями за класичною теорією ймовірності $c_j = 0{,}2$, при чотирьох запропонованих відповідях $c_j = 0{,}25$.

Сучасна теорія дозволяє отримати числові значення рівня досягнень випробуваного в логітах за інтервальною шкалою, що дозволяє використовувати потужний апарат математичної статистики для інтерпретації отриманих результатів .

Порівнюючи CRT та IRT, можна виділити вагомі переваги сучасної теорією:

а) стійкість і об'єктивність оцінок параметра, що характеризує рівень підготовки випробуваних. Джерелом стійкості є відносна інваріантність оцінок рівня підготовки від складності завдань;

б) можливість вимірювання значень параметрів випробуваних і завдань тесту за однією і тією ж шкалою, що має властивості інтервальної;

в) стійкість і об'єктивність оцінок параметра складності завдань, їх незалежність від властивостей вибірки учасників тесту.

У вітчизняній літературі найбільш повні методики оцінювання якості тесту в рамках класичної і сучасної теорій представлені в роботах В. Кіма [23] і С. Каракозова [24].

2.4 Автоматизовані системи обробки результатів на основі IRT

Застосування сучасної теорії тестів передбачає роботу з великими масивами даних, тому доцільніше використовувати автоматизовані системи для обробки результатів. Нижче подано короткий опис декількох систем, призначених для обробки результатів тестування у рамках деяких моделей IRT.

В компанії ASC розроблюються також програмні забезпечення, які використовують теорію IRT для обробки результатів тестувань.

LOGIST – програма для оцінки можливостей випробовуваного і характеристик кривої елементів. Перша версія якої була розроблена в 1976 р.

Недоліком даної програми є вимога великої вибірки випробовуваних більше ніж 1000 (для забезпечення більш точних оцінок) [21].

BICAL – програма розроблена в 1979 р. для калібрування параметрів однопараметричної моделі Г.Раша [25]. Оскільки BICAL не використовує алгоритм оцінки відсутніх даних то відсутні відповіді трактуються як неправильні. Ця програма надає різноманітну інформацію про калібрування для дихотомічних даних. Наприклад, вона може проводити повторне калібрування з видаленими невідповідними випробовуваними (ті, що набрали дуже низькі та дуже високі бали) із процесу калібрування. BICAL – одна з небагатьох програм, яка надає статистику відповідності між випробовуваним та показниками в групі, а також її результати дуже прості для інтерпретації [26].

BILOG – програмне забезпечення розроблене в 1984 р. дозволяє отримати оцінки параметрів тестових завдань на основі теорії IRT з використанням однопараметричної моделі Г. Раша, двопараметричної та трипараметричної моделі А. Бірнбаума для аналізу дихотомічних даних. BILOG дає змогу обробляти тести з максимум 1000 змінних на кожного випробуваного і без практичного обмеження на загальну кількість випробовуваних.

А також BILOG оцінює параметри елементів використовуючи оцінку граничної максимальної подібності [27], яка використовує EM-алгоритм (Expectation-maximization algorithm) розроблений А. Демпстером, Н. Лейрдом та Д. Рубіномтом у роботі «Maximum likelihood from incomplete data via the EM algorithm» [28].

FastTest – програмне забезпечення розроблене спеціально для використання професіоналами в галузі тестування і моделювання різних тестів, що підтримує режими бланкового і комп’ютерного пред’явлення паралельних варіантів тестів і оцінки їх якості за допомогою IRT моделей для дихотомічних даних за завданнями. FastTest акцентує увагу на оцінку якості тестування – валідність, надійність, безпеку, масштабованість, а також забезпечує вдосконалену психометрику з повною підтримкою теорії IRT, що дозволяє легко будувати адаптивні тести.

XCALIBRE, що дозволяє отримати оцінку найвищого правдивого підключення на основі алгоритмів EM для небольших вибірок досліджуваних або коротких тестів для дво- і трехпараметричних моделей IRT.

Окрім програм, що описані вище є ще інші MULTILOG (1991 р.), PARSCALE (1997 р.), WINMIRA (2001 р.) .

RUMM – програмний засіб розроблений під керівництвом Д .Эндрича (D. Andrich) в 1990 році для оцінки якості тестових завдань, який дозволяє оптимізувати зміст тесту і перетворювати його в інструмент для вимірювання рівня знань випробовуваних, тобто використовує методологію та технологію RM [29].

RUMM дозволяє реалізувати RM, лише тоді, коли дані відповідають моделі Г.Раша. Тобто, якщо результати тестового завдання не задовольняють моделі Г.Раша, то це завдання слід видалити з тесту, що не володіє вимірювальними властивостями. RUMM 2020 дає можливість аналізувати параметри тестових завдань відповідно до ймовірнісної моделі Г. Раша за спрощеними алгоритмами обчислень параметрів.

Алгоритм опрацювання результатів тестування програмою RUMM 2020, а також аналіз даних в політомічному випадку («0-1-2»)можна знайти у роботі Г. Смирнової [30].

Для аналізу результатів тестування в RUMM необхідно підготувати вхідний файл даних, що містить матрицю результатів тестування. Це звичайний текстовий файл з розширенням .dat, який можна створити за допомоги текстового редактора MS Word, «Блокнот», або ж у вигляді таблиці MS Excel. Як і в програмі Iteman 4, вхідний файл має фіксовану ширину та ділиться на два блоки: ідентифікатор та відповіді на завдання. Перший блок включає рядки, що містять номер досліджуваного і його прізвище. Ширина блоку обираємо за найдовшим прізвищем. Для всіх інших випробуваних з коротшим прізвищем відсутні символи заповнюються пробілами (рис. 2 .21):

```
«-1-Ivanov-I.------»
```

Рисунок 2.21 – Заповнення першого блоку у вхідному файлі програми RUMM

Символ «-» умовно показує порожні місця. Разом 18 символів. Таким чином, всі рядки першого блоку мають ширину 18 символів.

Другий блок містить відповідні рядки бінарної матриці. Його ширина дорівнює 10 символам, тобто кількості завдань у тесті (рис. 2.22).

```
1 Ivanov I.        1111111110
2 Sidorov A.       0101000000
3 Mikhaylov M.     1110110110
4 Alekseev D.      1011001011
5 Yakovlev A.      1111011010
6 Meleshko V.      1111110111
7 Petrov M.        1101101110
8 Krivchenko O.    1011011101
```

Рисунок 2.22 – Вхідний файл

Обчислювати ширину блоків не обов’язково – це зробить RUMM, потрібно забезпечити однакову ширину рядків в першому блоці. Кількість блоків у файлі може бути і більше двох – це залежить від ступеня деталізації опису експерименту.

Якщо дані представлені в форматі електронної таблиці MS Excel (рис. 2.23), необхідно виділити область бінарної матриці Бінарна матриця в Excel –це стовпці від «C» до «L» і встановити для них фіксовану ширину стовпчика рівну одиниці.

	A	B	C	D	E	F	G	H	I	J	K	L
1	1 Ivanov I.	1	1	1	1	1	1	1	1	1	0	
2	2 Sidorov A.	0	1	0	1	0	0	0	0	0	0	
3	3 Mikhaylov M.	1	1	1	0	1	1	0	1	1	0	
4	4 Alekseev D.	1	0	1	1	0	0	1	0	1	1	
5	5 Yakovlev A.	1	1	1	1	0	1	1	0	1	0	
6	6 Meleshko V.	1	1	1	1	1	1	0	1	1	1	
7	7 Petrov M.	1	1	0	1	1	0	1	1	1	0	

Рисунок 2.23 – Вхідний файл в форматі електронної таблиці

Створивши файлу вихідних даних можна запустити RUMM2020. На екрані з’явиться головне меню програми – «Main Menu» (рис. 2.24).

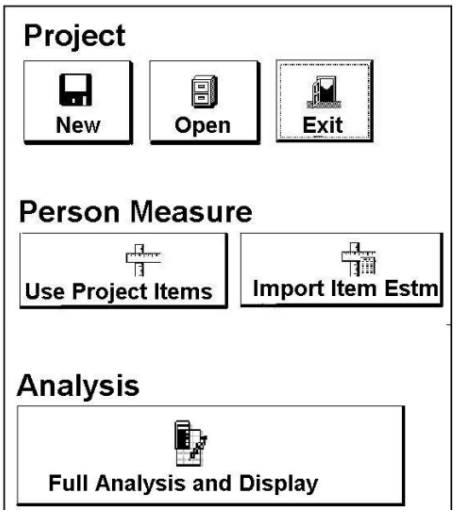


Рисунок 2.24 – Головне меню програми RUMM2020

Головне меню складається з трьох розділів: Project, Person Measure і Analisys. У кожному розділі є кілька кнопок, які дозволяють виконувати різні команди:

а) розділ Project:

1) New – створити новий проект (в форматі .mdb);

2) Open – відкрити існуючий проект;

3) Close – закрити поточний (використовуваний в даний момент) проект.

б) Person Measure:

1) Use Project Items – використовувати дані проекту;

2) Import Item Estm (estimates) – імпортувати (завантажити дані в іншому форматі);

в) Analisys – аналіз даних. Цей розділ з’являється, якщо вже відкривався файл проекту.

Команда Full Analisys and Display виконує аналіз даних і виведення результатів на дисплей.

У RUMM дані, специфікації аналізу і представлення даних зберігаються у вигляді файлів. Всі ці файли пов’язані воєдино в складі одного проекту. Тому, коли починається робота в RUMM, необхідно створити проект. Проект в RUMM є файлом у форматі бази даних Microsoft ACCESS тобто файл з розширенням .mdb.

Створивши проект за певними правилами програми RUMM, виконавши необхідні налаштування та назвавши його AN1, отримуємо діалогове вікно «ITEM CHARACTERISTIC CURVES (ICC) for Analysis Name AN1» (рис.2.24). На рис. 2.25 показаний графік 3-го завдання відповідно до моделі G.Rasch.

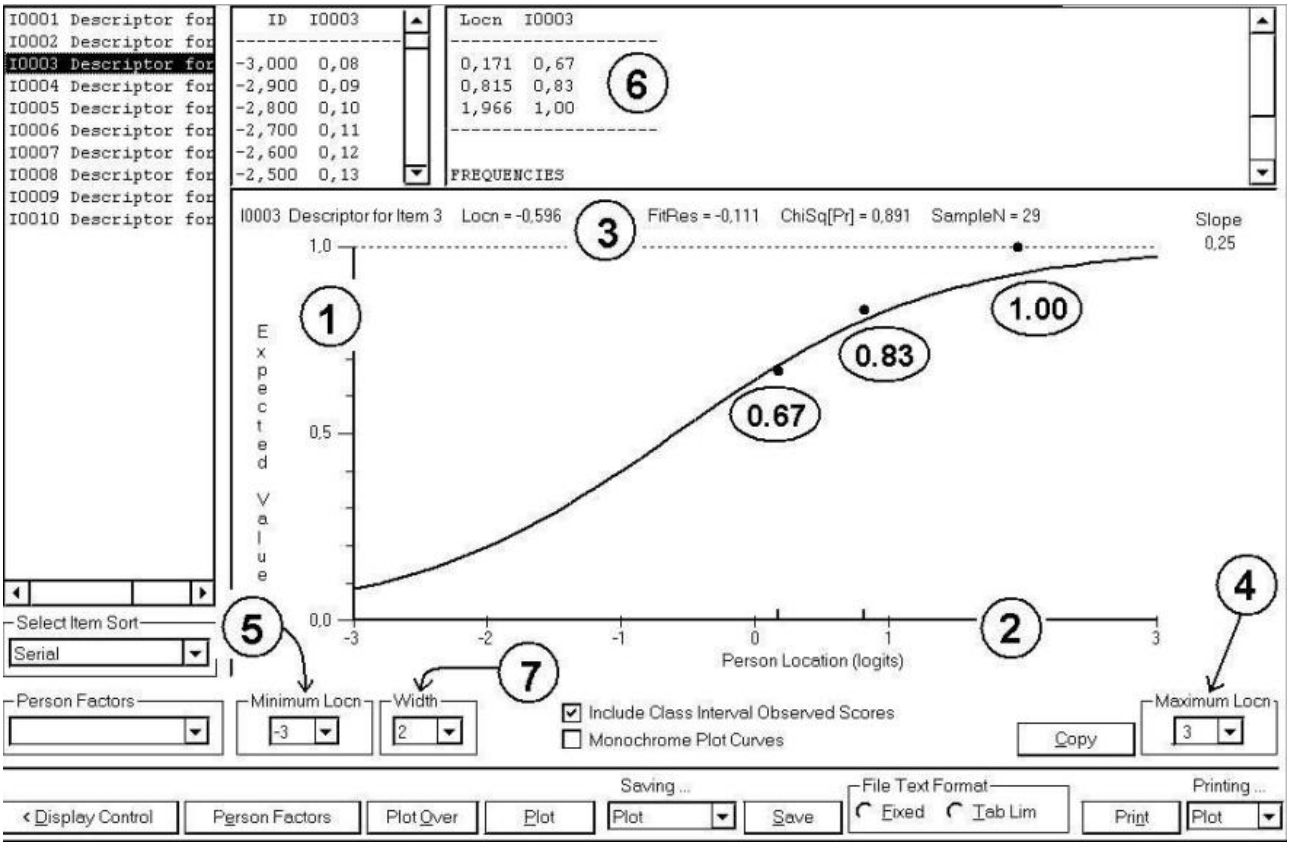


Рисунок 2.25 – Діалогове вікно «Item characteristic curves for Analysis Name AN1»

Цифрою «1» показана вісь ординат – «Expected Value» (Очікуване значення). Цифрою «2» – вісь абсцис – «Person Location (logits)». Цифрою «3» показана інформаційний рядок, що містить значення деяких характеристик другого тестового завдання (саме його дескриптор виділений в лівій частині вікна). Опис цих характеристик наведено в роботі Кіма В. С. [32].

Максимальне (Maximum Locn) мінімальне (Minimum Locn) значення параметра «Person Location» можна змінювати за допомогою випадаючих списків, показаних цифрами «4» і «5» відповідно. Параметр «Width», що позначений цифрою «7» визначає товщину лінії графіка.

На графіку наведені три точки, що відповідають експериментальним даним. Ці точки мають ординати 0.67, 0.83 і 1.00 відповідно. Координати цих точок наведені в таблиці, показаної цифрою «6». На графіку три точки тому, що при налаштуванні аналізу були встановлені три класових інтервалу (Class Intervals). До класових інтервалів групуються відносно близькості за результатами випробовувані. Для кожної групи розраховується одна емпірична точка [33].

Кількість класових інтервалів можна змінювати за допомогою списку, що показаного цифрою «1» на рис. 2.26.

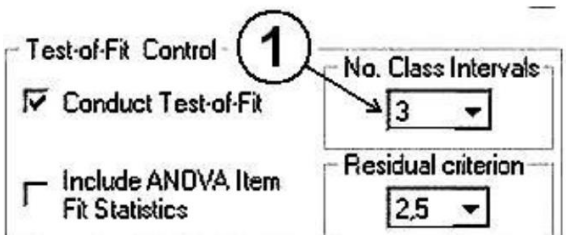


Рисунок 2.26 – Налаштування аналізу завдань у тесті

Після побудови характеристичної кривої завдань отримаємо результат, показаний на рис. 2.27.

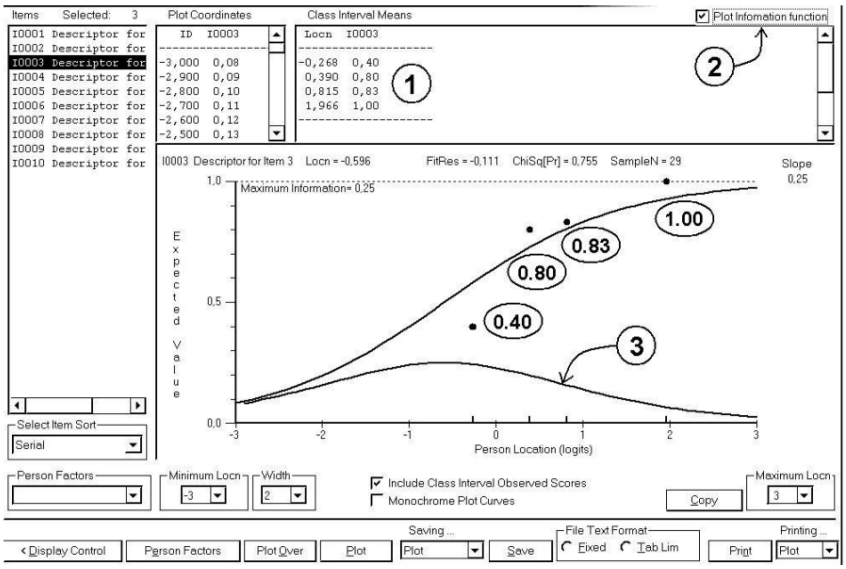


Рисунок 2.27 – Діалогове вікно «ICC for Analysis Name AN1»

Тепер таблиця «Class Interval Means» містить координати чотирьох точок (цифра «1»). Цифрою «2» показана команда «Plot Information function». Якщо команда включена, то одночасно з ICC будується інформаційна функція тесту, показана цифрою «3».

Отже, розглянувши методику побудови графіків завдань і графіків інформаційних функцій за допомогою програмного засобу RUMM 2020, можна зробити наступні висновки.

Підготовка вихідного файлу даних виконується різними способами і досить проста, але якщо великий масив даних, виникають незручності – велика витрата часу на введення даних. Дана програма працює з обробленими результати тестування, тобто вона не перевіряє початкові дані та правильність відповіді.

Інтерфейс системи досить простий, тому робота в RUMM не представляє особливих труднощів і вимагає тільки початкових навичок роботи на комп'ютері, а також знання англійської мови.

Популярною програмою є WINSTEPS або її Windows версія BIGSTEPS. WINSTEPS – програма 1991 року, яка дозволяє калібрувати дихотомічні завдання за допомогою моделі Раша та політомічні завдання у рамках моделей Partial Credit та Rating Scale, Paired Comparison, Success, Failure, а також різні комбінації цих моделей [34]. Для оцінки параметрів у програмі використовується процедура максимальної вірогідності JML.

Дана програма набула популярності частково завдяки своїй безкоштовній академічній версії MINISTEP, яка обмежена до 25 завдань та з вибіркою до 75 випробуваних [21]. Саме її будемо досліджувати.

Вхідні дані можна вносити безпосередньо у MINISTEP за допомогою пункту меню Data Setup, де забезпечено зручне автоматичне формування командних рядків робочого файлу (рис. 2.28). Стовпчики відповідають завданням, рядки – учасникам тестування. У кожній клітинці – кількість набраних балів (0 або 1 для дихотомічних завдань, від 0 до максимальної кількості балів за завдання – для політомічних).

Column:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Person:	1																							
Item No.	1																							
Label:																								
1	1	1	0	1	0	1	1	1	0	1	0	1	1	0	0	1	1	1	1	1	1	0	0	1
2	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
3	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
4	1	1	1	0	1	0	1	0	0	1	1	0	0	0	0	0	1	1	1	0	0	0	0	0
5	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
6	1	0	1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	1	1	1	1	1	0	0
7	1	1	0	0	0	0	0	1	0	1	0	1	0	0	1	1	0	0	0	1	0	0	1	1
8	1	1	1	1	1	1	1	0	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	1
9	1	1	1	0	1	1	1	0	1	0	1	0	1	0	1	1	1	1	0	0	1	0	0	1
10	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1
11	1	1	0	0	1	1	1	0	1	0	1	0	1	0	1	1	1	1	0	0	1	1	0	1

Рисунок 2.28 – Вхідна матриця результатів у програмі MINISTEP

Також програма працює з даними, які можна підготувати у різних статистичних пакетах, тому матрицю відповідей можна створити в таких програмах, як R, SAS, SPSS, MS Excel та ін. (рис. 2.29).

Рисунок 2.29 – Можливі форма матриць відповідей

Якщо використовуються дані з інших статистичних пакетів, необхідно провести конвертування даних за правилами програми [34]. Виконавши всі вимоги, переходимо до аналізу результатів.

Після початку аналізу програма виведе на екран загальні дані про кількість завдань та випробуваних, середні значення оцінених параметрів та дисперсію. Для більш докладнішого аналізу завдань та учасників потрібно скористатися пунктом головного меню Output Tables з різними табличними звітами або пунктом Graphs з графічними звітами(рис. 2.30).

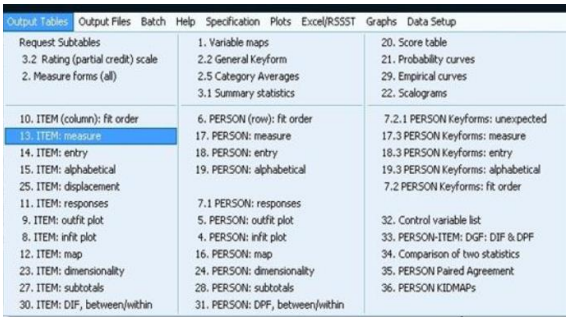


Рисунок 2.30 – Вкладка Output Tables у програмі MINISTEP

Результати конвертування вхідних даних у вимірювання Раша для всіх завдань тесту можна отримати, використавши команду Item: measure в меню Output Table (рис. 2.31).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
13	7	48	2.69	.44	1.40	1.3	1.54	1.2	-.01	.35	83.3	86.1	13
7	15	48	1.52	.34	.84	-1.0	.76	-1.1	.56	.39	79.2	74.5	7
2	17	48	1.30	.33	.95	-.3	.96	-.2	.44	.39	75.0	71.7	2
12	19	48	1.09	.32	1.20	1.6	1.40	2.1	.16	.39	62.5	69.1	12
4	25	48	.49	.31	.83	-1.8	.83	-1.1	.54	.38	77.1	64.3	4
10	25	48	.49	.31	.91	-.9	.83	-1.1	.48	.38	60.4	64.3	10
14	26	48	.39	.31	.89	-1.2	.82	-1.1	.50	.37	70.8	64.8	14
6	27	48	.30	.31	1.07	.7	1.05	.4	.30	.37	64.6	65.2	6
9	30	48	.00	.32	.92	-.7	.81	-.9	.46	.35	62.5	67.8	9
8	31	48	-.11	.32	1.06	.5	.97	.0	.31	.35	60.4	69.1	8
11	33	48	-.32	.33	1.16	1.2	1.32	1.3	.15	.33	66.7	71.8	11
5	43	48	-1.83	.49	.90	-.2	.70	-.3	.34	.22	89.6	89.6	5
1	45	48	-2.42	.61	.98	.1	.99	.3	.18	.18	93.8	93.8	1
3	47	48	-3.60	1.02	1.04	.4	.89	.3	.08	.11	97.9	97.9	3
MEAN	27.9	48.0	.00	.41	1.01	.0	.99	.0			74.6	75.0	
S.D.	11.2	.0	1.59	.19	.15	1.0	.25	1.0			12.3	11.3	

Рисунок 2.31 – Статистичні дані завдань тесту

У першому стовпчику (ENTRY NUMBER) вказані номери завдань, у другому (TOTAL SCORE) – кількість правильних відповідей даних на це питання, у третьому (TOTAL COUNT) – загальна кількість усіх відповідей. Результати вимірювання складності завдань у логітах показано у порядку спадання в четвертому стовпчику (MEASURE). У стовпчику MODEL S.E. наведена похибка вимірювання на основі моделі Раша, а у рядках MEAN та S.D. – середні значення та стандартні відхилення для значень у відповідних стовпчиках [34].

У стовпчиках INFIT та OUTFIT знаходяться параметри, що характеризують відповідність даних моделі Раша. Значення MNSQ (mean-square statistic) характеризують рівень випадковості результатів або невідповідність даних моделі вимірювання. Найбільш очікувані значення MNSQ знаходяться поблизу 1. Великі значення MNSQ OUTFIT пов’язують з угадуванням відповідей, а великі MNSQ INFIT інтерпретуються як показник низької валідності завдань [21]. Значення MNSQ більші за значення 2 розглядаються як такі, що не відповідають моделі вимірювання і не можуть бути використані при аналізі результатів. Найбільш вдалими вважаються значення MNSQ у межах від 0,5 до 1,5. Якщо значення більші за 1,5 – це вказує на невизначеність та «шум» у

вхідних даних, якщо ж значення менші за 0,5 – це свідчить про «інформаційну перевантаженість» питання.

Аналіз починають із питань з високим значенням MNSQ. У полі ZSTD наводяться стандартизовані значення MNSQ. Прийнятними є значення від –2 до 2. Стовпчик PT-MEASURE CORR. Відповідає за значення коефіцієнта кореляції, який змінюється від –1 до +1. Він розглядається як деякий показник надійності та валідності та може бути використаний для визначення, доопрацювання, а можливо і виключення слабко узгоджених завдань.

Стовпчик OBS% – відсоток балів даних, отриманих при спостережені, які знаходяться в межах значення 0,5 від їх очікуваних значень, тобто, що відповідають прогнозам. Стовпчик EXP% – очікуваний відсоток балів даних, які, за прогнозами, будуть у межах значення 0,5 від їх очікуваних значень.

Якщо OBS% <EXP%, то локальні дані більш випадкові, ніж прогнозує модель Раша. Якщо OBS%> EXP%, то локальні дані є більш передбачуваними, ніж прогнозує модель.

Також можна також отримати звіт з характеристиками учасників тестування командою Person: measure в меню Output Tables (рис. 2.32).

ENTRY NUMBER	TOTAL SCORE	COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
6	12	14	2.54	.98	1.01	.3	1.02	.5	.52	.54	92.9	91.3	P6
7	12	14	2.54	.98	1.01	.3	1.02	.5	.52	.54	92.9	91.3	P7
11	12	14	2.54	.98	1.01	.3	1.02	.5	.52	.54	92.9	91.3	P11
26	6	14	-.44	.63	1.07	.4	.84	.2	.47	.48	50.0	71.6	P26
27	6	14	-.44	.63	1.22	1.0	9.90	4.2	.17	.48	64.3	71.6	P27
36	6	14	-.44	.63	.93	-.2	.73	.0	.53	.48	64.3	71.6	P36
3	4	14	-1.30	.70	1.03	.2	.76	.2	.48	.47	71.4	79.8	P3
4	4	14	-1.30	.70	1.49	1.3	1.78	.9	.24	.47	71.4	79.8	P4
29	4	14	-1.30	.70	1.03	.2	.76	.2	.48	.47	71.4	79.8	P29
20	3	14	-1.85	.78	1.06	.3	.82	.2	.46	.47	78.6	84.5	P20
18	2	14	-2.57	.93	.96	.2	1.03	.4	.44	.45	92.9	89.9	P18
MEAN	7.9	14.0	.39	.70	1.00	.0	1.13	.3			74.4	76.9	
S.D.	2.5	.0	1.18	.12	.20	.7	1.47	.7			13.6	7.3	

Рисунок 2.32 – Статистичні дані учасників тестування

У MINISTEP можна отримати різні графічні звіти: характеристичні криві, інформаційні функції тощо. На рис. 2.33 побудовані характеристичні криві усіх завдань деякого дихотомічного тесту, аналіз взаємного розміщення яких допомагає вдосконалити тест як систему завдань зростаючої складності.

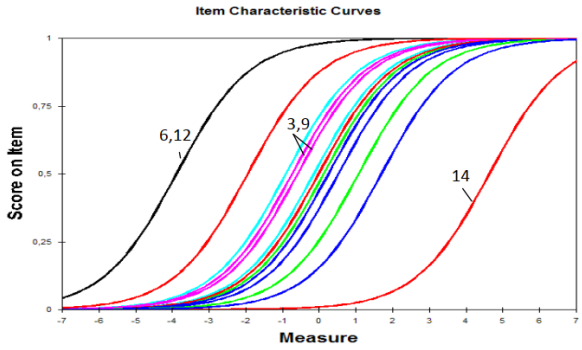
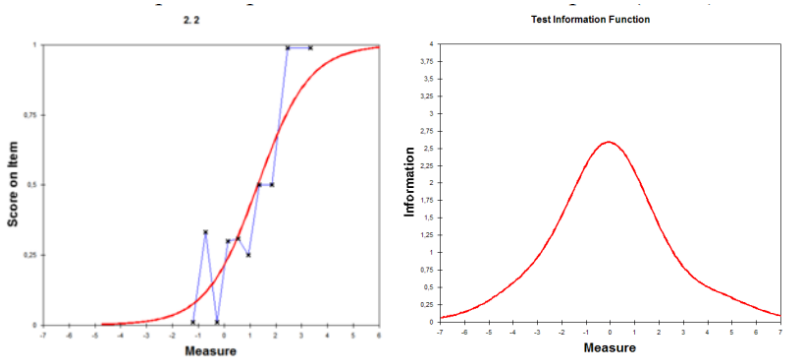


Рисунок 2.33 – Графік характеристичних кривих завдань тесту

Якщо тест був правильно створений, то криві повинні бути представлені рівномірно на всьому інтервалі (–5;+5), без накладувань та характеристичні криві завдань розташовані в порядку зростання. У разі накладань кривих одне з них можна вилучити. Такі завдання можуть бути використані для паралельних тестів. Для кожного завдання та тесту загалом можна отримати графічне представлення відповідності даних обраній моделі (рис. 2.33, а) та інформаційні функції (рис. 2.34, б).



а) б)

Рисунок 2.34 – Графік відповідності даних моделі (а), графік інформаційної функції (б)

Крім того, за допомогою пункту меню Plots можна отримати табличне та графічне представлення результатів в Excel. Використовуючи пункт Output Files, можна зберегти вихідні файли в SPSS або в R-Statistics.

Проаналізувавши основні можливості програми MINISTEP, можна виділити переваги:

а) простий та зручний у використані інтерфейс;

б) працювати з даними, які були створені в інших статистичних програм;

в) отримані вихідні дані зберігати в інших форматах;

г) самостійно обирати статистичні процеси.

До недоліків програми можна віднести, що на обробку результатів тестування можна винести лише 25 завдань та 75 тестованих.

2.5 Висновок до другого розділу

Таким чином автоматизовані системи обробки результатів тестування Iteman 4, Lertap, CITAS. MyTestXPro, R, SPSS Statistics використовують класичну теорію обробки результатів тестування, а системи LOGIST, BICAL, BILOG, FastTest, XCALIBRE, MULTILOG, PARSCALE, WINMIRA, RUMM, WINSTEPS засновані на IRT теорії.

Програми Iteman 4, MyTestXPro, RUMM, WINSTEPS були детально розглянуті у даному розділі, було виділено основні можливості систем, оцінка зручності інтерфейсу, преваги та їх недоліки. Результати цього аналізу подано у порівняльній таблиці 2.5.

Таблиця 2.5 – Порівняльний аналіз автоматизованих систем обробки результатів тестування

Категорії порівняння	Iteman (безкоштовна обмежена версія)	Ministep (безкоштовна академічна версія)	Rumm
Введення вхідних даних у програму	+	+	+
Конвертування вхідного файлу		+	
Конвертування вихідного файлу		+	

Продовження таблиці 2.5

Категорії порівняння	Iteman (безкоштовна обмежена версія)	Ministep (безкоштовна академічна версія)	Rumm
Дослідження правдоподібності дистракторів	+		
Використання моделі Раша		+	+
Обмеження розмірності на вхідні дані	100×100	25×75	
Дослідження ефективності тесту		+	+
Обчислення коефіцієнтів надійності по завданням	+	+	+
Обчислення коефіцієнтів надійності по випробуваним	+	+	+
Валідність тесту	+	+	+

3 ОБРОБКА РЕЗУЛЬТАТІВ ТЕСТУВАННЯ АВТОМАТИЗОВАНИМИ СИСТЕМАМИ ITEMAN, MINISTEP ТА ЇХ ПОРІВНЯЛЬНИЙ АНАЛІЗ

Для проведення порівняльного аналізу ефективностей систем Iteman 4 (безкоштовна обмежена версія) та Ministep (безкоштовна академічна версія) оброблено результати тестування за допомогою програми MS Excell, використовуючи CRT та IRT теорії.

3.1 Обробка результатів тестування пробного ЗНО за допомогою програми MS Excell

Для дослідження якості тесту було обрано результати тестування пробного ЗНО з математики лише тестової частини, тобто 20 завдань, що проводилось у навчальному закладі ЗНУ. Вибірка з 50 випробуваних. Досліджування результатів за CRT теорією проводилось за алгоритмом запропонованим Авраменко О. В. [17].

Обробка та аналіз результатів тестування поділяється на декілька етапів:

а) первинний аналіз результатів тестування:

1) редукувати матриця результатів тестування (порахувати індивідуальні бали учасників тестування, порахувати кількість правильних відповідей для кожного із тестових завдань, упорядкувати бінарну матрицю, вилучити з матриці рядки та стовпці, яку не дають жодної інформації про тест та випробуваних);

2) побудувати ряди результатів тестування та графічно інтерпретувати їх (побудувати частотний ряд, побудувати гістограму частот).

б) основні статистичні характеристики результатів тестування:

1) знайти міри центральної тенденції тестових балів (моду, медіану, середнє вибіркове);

2) знайти міри мінливості тестових балів (розмах, дисперсію, стандартне відхилення);

3) перевірити гіпотезу про нормальний закон розподілу результатів тестування;

4) обчислити асиметрію та ексцес.

в) кореляційний аналіз результатів тестування:

1) побудувати кореляційну матрицю тестових завдань;

2) обчислити точково-бісеріальні коефіцієнти кореляції.

г) основні характеристики тестових завдань:

1) оцінити трудність тестових завдань;

2) визначити рівень правдоподібності дистракторів.

Спочатку будуємо первинну матицю результатів, рахуємо індивідуальні бали учасників та кількість правильних відповідей для кожного із тестових завдань (рис. 3.1).

	Номер завдання																			
Номер студ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Xi
студ 1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	1	1	0	1	0	0	7
студ 2	1	1	0	0	1	1	1	1	1	0	0	1	1	0	1	0	1	1	0	1	13
студ 3	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	17
студ 4	1	0	0	0	1	1	1	1	0	0	0	1	0	0	1	1	0	1	0	1	10
студ 5	1	0	0	0	1	1	1	1	0	0	0	1	0	0	1	0	0	1	0	1	9
студ 6	1	1	1	1	0	1	1	1	1	0	0	1	1	0	1	1	1	1	1	0	15
студ 7	0	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	1	1	9
студ 8	1	1	0	0	1	1	1	0	0	0	1	0	0	0	0	1	0	0	0	0	7
студ 9	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0	1	0	4
студ 10	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0	1	0	14
студ 11	1	1	1	0	1	1	1	0	0	0	0	0	0	1	0	0	0	0	0	1	8
студ 12	0	1	1	0	1	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	8
студ 13	0	1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	8
студ 14	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1	0	15
студ 15	1	1	1	1	0	1	1	0	1	0	0	0	0	1	0	0	0	0	1	1	10
студ 16	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1	0	15
студ 17	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	1	0	1	8
студ 18	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1	15
студ 19	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1	15
студ 20	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	18
студ 21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20

Рисунок 3.1 – Матриця первинних результатів тестування

Після упорядковування бінарної матриці (рис. 3.2), отримуємо, що студент 21 відповів на всі завдання вірно, а отже його потрібно вилучити з матриці, тому що даний тест є непридатний для оцінки якості знань цього студента. Причина непридатності тесту – надмірна легкість, що не дозволяє виявити справжній рівень його знань.

Номер студ	Номер завдання																				Xi
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
студ 21	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	20
студ 22	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	19
студ 26	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19
студ 20	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	18
студ 25	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	18
студ 3	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	17
студ 6	1	1	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	17
студ 30	1	1	1	1	1	0	1	1	1	0	0	1	1	1	0	1	1	1	1	1	16
студ 14	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1	0	15
студ 16	1	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	0	1	0	15
студ 18	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1	15
студ 19	1	1	1	1	0	1	1	1	1	0	0	1	0	1	0	1	1	1	1	1	15
студ 43	1	1	1	1	1	1	1	0	1	0	1	1	1	1	0	1	0	1	1	0	15
студ 10	1	1	1	1	1	1	1	1	1	0	0	1	1	0	0	1	1	0	1	0	14
студ 23	1	1	0	0	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	14
студ 28	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	1	1	0	0	1	14
студ 29	1	1	1	0	1	1	1	0	1	0	1	1	1	0	1	0	1	1	1	0	14
студ 2	1	1	0	0	1	1	1	1	1	0	0	1	1	0	1	0	1	1	0	1	13
студ 24	1	1	0	0	1	1	1	1	0	0	0	1	0	1	1	1	1	0	1	1	13
студ 27	1	1	0	1	1	1	1	0	1	0	0	1	1	1	0	0	0	1	1	1	13
студ 50	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	0	1	13

Рисунок 3.2 – Редукована матриця результатів

Упорядкувавши нумерацію завдань за кількістю правильних відповідей та присвоївши нові номери (рис. 3.3), видно, що у тесті порушено диференційованість завдань.

Старі номери	1	7	2	6	5	12	18	9	16	4	3	17	19	20	14	8	15	13	10	11
Нові номери	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Рисунок 3.3 – Нова нумерація завдань

Надалі будемо працювати з новою нумерацією завдань та випробуваних.

Побудувавши частотний ряд індивідуальних балів *Xi*, зображуємо відповідну гістограму (рис. 3.4).

Рисунок 3.4 – Гістограма частот індивідуальних балів

За класичною теорією, крива частот індивідуальних балів якісного тесту повинна відповідати нормальному закону розподілу. Аналізуючи графік гістограми, гіпотезу про нормальний розподіл результатів тестування для даного тесту ми не можемо прийняти.

Виконавши пункт б) з обробки результатів тестування, отримано такі дані:

а) мода = 8 та 15;

б) медіана = 11;

в) середнє вибіркове = 11,122;

г) розмах = 17;

д) дисперсія = 17,026;

е) стандартне відхилення = 4,126;

є) асиметрія = 0,028;

ж) ексцес = -0,648.

Перевірено гіпотезу про нормальний закон розподілу результатів тестування за критерієм Пірсорна та отримано $К_{емпір}$=15,848, а $К_{кр}$=27,249 при α=0,005.

За допомогою вище згаданих даних, можна зробити висновок, що гіпотезу про нормальний закон розподілу результатів тестування за критерієм Пірсорна не відхилено. Нормальний закон унімодальний і симетричний, а мода, медіана та середнє значення повинні бути рівні. Для нашого випадку, значення двох останніх параметрів приблизно однакові, що є гарним показником, але моди дві. Значення дисперсії завелике, що говорить про не рівномірну диференціацію випробуваних у групі. Значення асиметрії близьке до нуля, тобто тест добре збалансований за трудністю завдань. А від’ємне значення ексцесу означає, що результати тестування є сильно розкиданими в околі їх середнього значення, як показано на гістограмі вище.

Провівши кореляційний аналіз результатів тестування, отримано кореляційну матрицю тестових завдань (рис. 3.5) та обчислено бісеріальні коефіцієнті кореляції для кожного тестового завдання (рис. 3.6).

	Завд 1	Завд 2	Завд 3	Завд 4	Завд 5	Завд 6	Завд 7	Завд 8	Завд 9	Завд 10	Завд 11	Завд 12	Завд 13	Завд 14	Завд 15	Завд 16	Завд 17	Завд 18	Завд 19	Завд 20
Завд 1	1,000	0,402	0,262	0,360	-0,157	0,465	0,236	0,307	-0,048	0,267	0,156	0,045	0,156	0,101	0,288	0,254	0,238	0,108	-0,110	0,252
Завд 2	0,402	1,000	0,134	0,485	0,087	0,351	0,122	0,083	0,064	0,267	0,267	-0,066	0,267	0,212	0,288	0,367	0,238	0,108	0,140	0,252
Завд 3	0,262	0,134	1,000	0,303	-0,045	-0,040	-0,139	0,203	-0,205	0,154	0,345	0,154	0,249	-0,107	0,323	0,183	-0,131	0,139	-0,088	0,214
Завд 4	0,360	0,485	0,303	1,000	0,029	0,213	0,213	0,253	-0,147	0,108	0,108	0,108	0,201	0,034	0,264	0,311	0,099	0,074	0,047	0,235
Завд 5	-0,157	0,087	-0,045	0,029	1,000	-0,107	-0,013	-0,066	0,274	-0,208	-0,208	0,156	0,247	-0,013	0,026	0,066	0,132	0,295	-0,029	0,045
Завд 6	0,465	0,351	-0,040	0,213	-0,107	1,000	0,297	0,229	0,196	0,248	0,078	0,163	0,163	0,238	0,007	0,547	0,172	0,405	0,266	0,139
Завд 7	0,236	0,122	-0,139	0,213	-0,013	0,297	1,000	0,142	0,196	0,163	-0,092	0,163	0,163	0,069	0,262	0,116	0,259	-0,034	-0,117	0,139
Завд 8	0,307	0,083	0,203	0,253	-0,066	0,229	0,142	1,000	0,036	0,586	0,336	0,336	0,252	-0,017	-0,002	0,183	-0,106	0,288	0,123	0,087
Завд 9	-0,048	0,064	-0,205	-0,147	0,274	0,196	0,196	0,036	1,000	0,213	0,130	0,213	0,296	0,106	0,036	0,384	-0,073	0,232	0,053	0,014
Завд 10	0,267	0,267	0,154	0,108	-0,208	0,248	0,163	0,586	0,213	1,000	0,505	0,175	0,340	0,228	0,237	0,249	-0,040	0,177	0,264	0,037
Завд 11	0,156	0,267	0,345	0,108	-0,208	0,078	-0,092	0,336	0,130	0,505	1,000	0,093	0,258	0,064	0,072	0,332	-0,124	0,262	0,264	0,132
Завд 12	0,045	-0,066	0,154	0,108	0,156	0,163	0,163	0,336	0,213	0,175	0,093	1,000	0,423	-0,018	0,155	0,332	0,129	0,177	-0,015	0,037
Завд 13	0,156	0,267	0,249	0,201	0,247	0,163	0,163	0,252	0,296	0,340	0,258	0,423	1,000	0,064	0,320	0,332	0,129	0,347	-0,108	-0,058
Завд 14	0,101	0,212	-0,107	0,034	-0,013	0,238	0,069	-0,017	0,106	0,228	0,064	-0,018	0,064	1,000	0,347	0,432	0,226	0,016	-0,034	-0,083
Завд 15	0,288	0,288	0,323	0,264	0,026	0,007	0,262	-0,002	0,036	0,237	0,072	0,155	0,320	0,347	1,000	0,085	0,208	-0,092	0,015	0,154
Завд 16	0,254	0,367	0,183	0,311	0,066	0,547	0,116	0,183	0,384	0,249	0,332	0,332	0,332	0,432	0,085	1,000	0,191	0,401	0,159	0,010
Завд 17	0,238	0,238	-0,131	0,099	0,132	0,172	0,259	-0,106	-0,073	-0,040	-0,124	0,129	0,129	0,226	0,208	0,191	1,000	-0,085	0,186	0,131
Завд 18	0,108	0,108	0,139	0,074	0,295	0,405	-0,034	0,288	0,232	0,177	0,262	0,177	0,347	0,016	-0,092	0,401	-0,085	1,000	0,117	0,058
Завд 19	-0,110	0,140	-0,088	0,047	-0,029	0,266	-0,117	0,123	0,053	0,264	0,264	-0,015	-0,108	-0,034	0,015	0,159	0,186	0,117	1,000	0,195
Завд 20	0,252	0,252	0,214	0,235	0,045	0,139	0,139	0,087	0,014	0,037	0,132	0,037	-0,058	-0,083	0,154	0,010	0,131	0,058	0,195	1,000
Сума	4,584	5,067	2,907	4,299	1,511	5,030	3,146	4,253	2,969	4,972	3,979	3,759	5,041	2,864	3,993	5,934	2,780	3,993	2,329	2,988

Рисунок 3.5 – Кореляційна матриця тестових завдань

Завдання пробного ЗНО перевіряють матеріал математики 7-11 класів, тому кореляція завдань одне з одним не повинна бути дуже високою, тобто не перевищувати 0,3. Аналізуючи значення кореляційної матриці, можна виділити завдання 5, 14, 17 та 19, які від'ємно корелюють з більшістю завдань. Ці завдання можуть мати помилки або відсутня предметна чистота змісту завдання. Тому слід переглянути завдання на помилки та коректність поставленого запитання у завданні. Можна виділити пари завдавань 2 та 4, 6 та 16, 8 та 10, 10 та 11, у яких кореляційні значення між собою більші за 0,3. Ці завдання можуть перевіряти знання з однієї теми.

Завдання	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Коеф бісер корел	0,500	0,554	0,391	0,471	0,248	0,572	0,365	0,493	0,359	0,580	0,459	0,449	0,590	0,340	0,455	0,687	0,239	0,464	0,265	0,211

Рисунок 3.6 – Бісеріальні коефіцієнті кореляції завдань

Точково-бісеріальний коефіцієнт кореляції використовується для оцінювання валідності окремих завдань. Загалом завдання можна вважати валідним, коли значення даного коефіцієнта близьке до 0,5. Оцінка валідності завдання дозволяє показати, наскільки воно придатне для роботи у відповідності з загальною метою тесту. Мета ЗНО – диференціація випробуваних за рівнем підготовки, тому валідні завдання повинні чітко відокремити добре підготовлених від слабо підготовлених учнів.

Проаналізувавши значення точково-бісеріальних коефіцієнтів кореляції тестових завдань з усім тестом пробного ЗНО, можна виділити завдання 5,17, 19 та 20, які мають невисоку корельованість. Їх структуру та зміст необхідно проаналізувати експертам. Усі інші завдання тесту мають значення кореляції у прийнятному інтервалі. Середнє значення кореляції завдань тесту – 0,435.

Переходимо до оцінки трудності (за термінологією О. В. Авраменка) тестових завдань. Потрібно зауважити, що за класичною теорією трудність завдань тим більше, чим більше учасників тестування його розв'язали. У IRT теорії тестування ця некоректність виправлена. Обчисливши трудність кожного завдання у тесті та його дисперсію, отримано дані, що зображено на рис. 3.7.

Номер завд	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
pj	0,837	0,837	0,755	0,735	0,714	0,633	0,633	0,592	0,571	0,551	0,551	0,551	0,551	0,490	0,449	0,408	0,388	0,367	0,265	0,245
qj	0,163	0,163	0,245	0,265	0,286	0,367	0,367	0,408	0,429	0,449	0,449	0,449	0,449	0,510	0,551	0,592	0,612	0,633	0,735	0,755
Дисп завд	0,137	0,137	0,185	0,195	0,204	0,232	0,232	0,242	0,245	0,247	0,247	0,247	0,247	0,250	0,247	0,242	0,237	0,232	0,195	0,185

Рисунок 3.7 – Дисперсія завдань тесту та трудність завдань

Так, як ми користуємось для дослідження новою нумерацією, то значення трудності для кожного завдання у тесті розміщені у порядку спадання. Крайні завдання 1 – 4, 19, 20 мають невелику дисперсію. Такі завдання рекомендовано включати у невеликій кількості у збалансований тест, як і в нашому тесті. Завдання 10-15 роблять максимальний внесок у загальну дисперсію тесту. Вони знаходяться в центральній частині ряду, що є гарним показником для тесту. Середній рівень трудності завдань нашого тесту – 0,556. Дане значення близьке до 0,5, а це є вдалим результатом.

Досліджено рівень правдоподібності дистракторів завдання 1 (рис. 3.8).

Відповіді	1	2	вірно	4	5
Prop.	0,04	0,02	0,82	0,06	0,04
Rpbis	-0,406	-0,109	0,520	-0,214	-0,181

Рисунок 3.8 – Рівень правдоподібності дистракторів завдання 1

За умови вдало обраних дистракторів слід очікувати рівномірний розподіл неправильних відповідей [17]. З отриманих даних, видно, що дистрактори для

завдання 1 підібрані не коректно, 0,82 доля випробуваних обрала вірну відповідь. Для поглибленого аналізу обчислено точко-бісеріальний коефіцієнт кореляції для кожного дистрактора в завдані. Якщо коефіцієнт бісеріальної кореляції дистрактора від'ємний і менший за -0,2, то сильні випробувані не будуть обирати його. Додатні та близькі до нуля значення коефіцієнта вказують на необхідність їх вилучення для вдосконалення неправильних відповідей. Значення коефіцієнта кореляції правильної відповіді повинно перевищувати 0,5. Для завдання 1 дистрактори 2 та 5 мають близькі значення до нуля, тому їх необхідно замінити. Коефіцієнт кореляції для вірної відповіді – 0,52, що є гарним показником.

За IRT теорією тестування можна побудувати характеристичні криві для завдань тесту однопараметричної моделі Раша (рис. 3.9), тобто встановити ймовірність виконання завдання, знаючи рівень підготовки випробуваного та рівень складності. Розрахунки виконано згідно М. Челишкої [16].

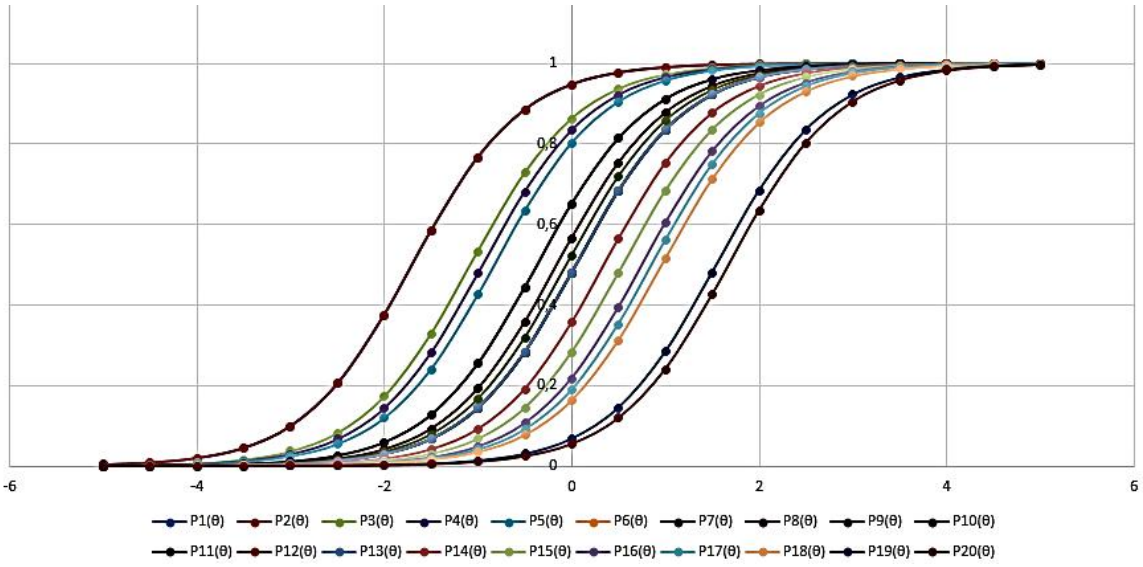


Рисунок 3.9 – Характеристичні криві для завдань тесту

Для диференційованого тесту характеристичні криві розташовані рівномірно, не перетинаються і не накладаються одну на одну. Тому пари завдань 1 та 2, 6 та 7, а також завдання 10 – 13 необхідно переглянути, бо їх характеристичні криві лежать одна на одній.

Проаналізувавши всі отримані дані, можна зробити рекомендації щодо покращення тесту:

а) змінити порядок завдань;

б) переглянути завдання 5, 14, 17 та 19 на помилки та коректність поставленого запитання у завданні;

в) переглянути пари завдавань 2 та 4, 6 та 16, 8 та 10, 10 та 11, які можливо перевіряють знання з однієї теми;

г) ускладнити завдання, щоб характеристичні криві завдань розміщувались рівномірно та не накладались одна на одну.

3.2 Обробка результатів тестування пробного ЗНО за допомогою програми Iteman

Ті ж самі результати оброблено за допомогою програми Iteman. Для використання цієї програми створено два файли – вхідний дані та файл контроль, в програмі MS Excell у форматі CSV. Після запуску системи отримано три звіти, у роботі подано повний первинний результат у форматі DOCX у Додатку 1.

У звіті спочатку подано таблиця, в якій представлена основна інформація щодо аналізу (рис. 3.10).

Specification¤	Value¤	Specification¤	Value¤
Number·of·examinees¤	50¤	Total·Items¤	20¤
Scored·Items¤	20¤	Pretest·Items¤	¤
Multiple·Choice·Items¤	20¤	Polytomous·Items¤	0¤
Number·of·Domains¤	1¤	External·scores¤	No¤
Minimum·P¤	-0,00¤	Maximum·P¤	1,00¤
Minimum·item·mean¤	0,00¤	Maximum·item·mean¤	15,00¤
Minimum·item·correlation¤	-0,00¤	Maximum·item·correlation¤	1,00¤
ITEMAN·3.0·Header¤	No¤	Exclude·omits·from·option·statistics¤	No¤
Number·of·ID·columns¤	6¤	ID·begins·in·column¤	1¤
Responses·begin·in·column¤	7¤	Omit·character¤	O¤
Not·Admin·character¤	N¤	Produce·quantile·tables¤	Yes¤
Correct·for·spuriousness¤	Yes¤	Produce·quantile·plots¤	Yes¤
Save·data·matrix¤	No¤	Include·omit·codes·in·matrix¤	N/A¤
Include·Not·Admin·codes·in·matrix¤	N/A¤	Include·scaled·scores·for¤	N/A¤

Рисунок 3.10 – Основна інформація аналізу

У таблиці 3.1 представлені основні статистичні характеристики щодо завдань тесту.

Таблиця 3.1 – Статистичні характеристики завдань у тесті

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
Scored Items	20	11,300	4,273	2	20	0,565	0,363

З даних таблиці 3.1 видно, що результати студента 21, який відповів на всі завдання вірно, були також враховані в статистичні обчислення. Тому табличні значення та значення, отримані за допомогою програми MS Excell відрізняються на незначну частину:

а) середнє вибіркове = 11,3;

б) стандартне відхилення = 4,273;

в) середнє значення складності завдань у тесті = 0,565;

г) середнє значення кореляції завдань у тесту = 0,363.

У таблиці 3.2 представлений аналіз надійності тестів. Альфа є найбільш часто використовуваним показником надійності, і тому використовується для обчислення стандартної похибки вимірювання (SEM) за необробленою шкалою. Також представлені три конфігурації надійності з розділеною половиною, спочатку як некоректовані кореляції, а потім як корекції кореляцій Спірмена-Брауна (S-B). Це пояснюється тим, що некорекційне співвідношення розділеної половини посилається на «тест», який містить лише половину досліджуваного тесту, і тому недооцінює надійність.

Таблиця 3.2 – Аналіз надійності тесту

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0,800	1,912	0,740	0,587	0,663	0,851	0,740	0,797

Оцінюючи всі значенні надійності, можна вважати даний тест надійним.

Також у звіті показано гістограму частот індивідуальних балів випробуваних (рис. 3.11) та відповідно їй таблицю.

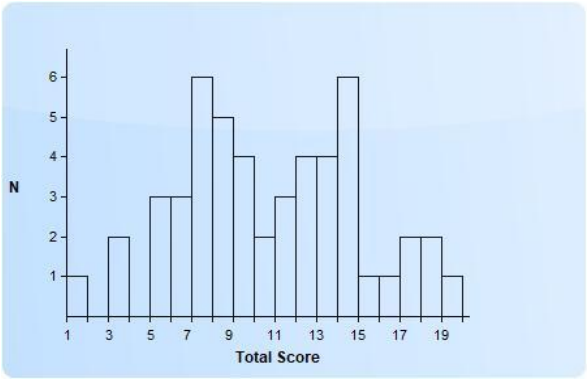


Рисунок 3.11 – Гістограма частот індивідуальних балів

Дана гістограма точно така, як і гістограма з попереднього пункту, лише з додаванням результату студента 21.

Наступним пунктом у звіті, графічно зображено кількість завдань, що мають певній рівень трудності (рис. 3.12).

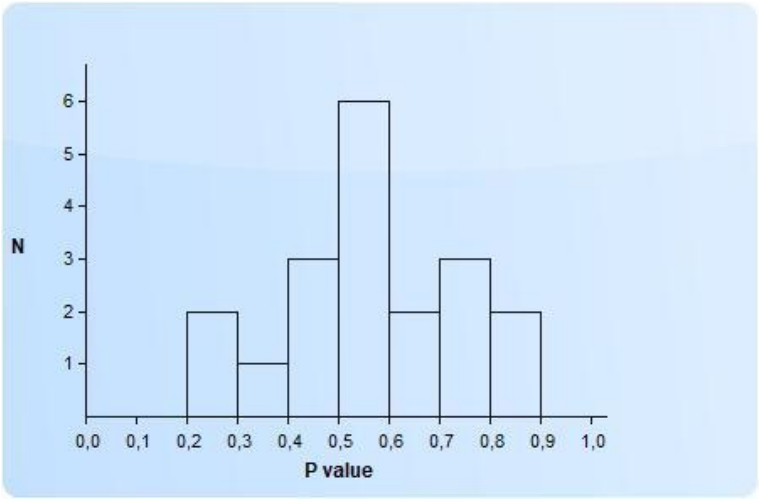


Рисунок 3.12 – Гістограма трудності завдань тесту

За допомогою гістограми видно, що найбільша кількість завдань має рівень трудності в інтервалі 0,5-0,6, що є гарним показником для тесту. Але для покращення якості тесту, рекомендовано збільшити або зменшити кількість завдань з певним рівнем трудності так, щоб гістограма була симетрична відносно інтервалу 0,5-0,6, як піраміда.

Також подано гістограму міру дискримінації завдань у тесті за допомогою точково-бісеріального коефіцієнта кореляції (рис. 3.13) та таблицю, що зображена на графіку.

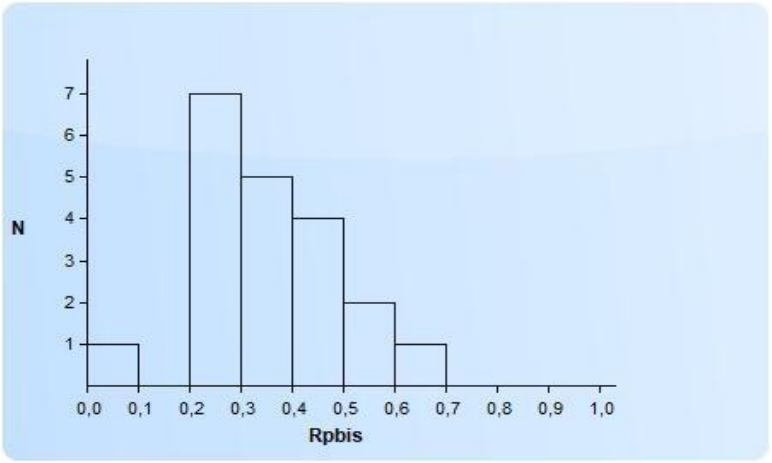


Рисунок 3.13 – Міра дискримінації завдань у тесті

З графіку видно, більшість завдань мають низькі значення точково-бісеріального коефіцієнта кореляції. Завдання, які мають значення Rpbis<0,3, погано диференціюють випробуваних. Необхідно переглянути завдання.

Після цього, показано розсіювання P (складність) і Rpbis (дискримінація) завдань у тесті (рис. 3.14).

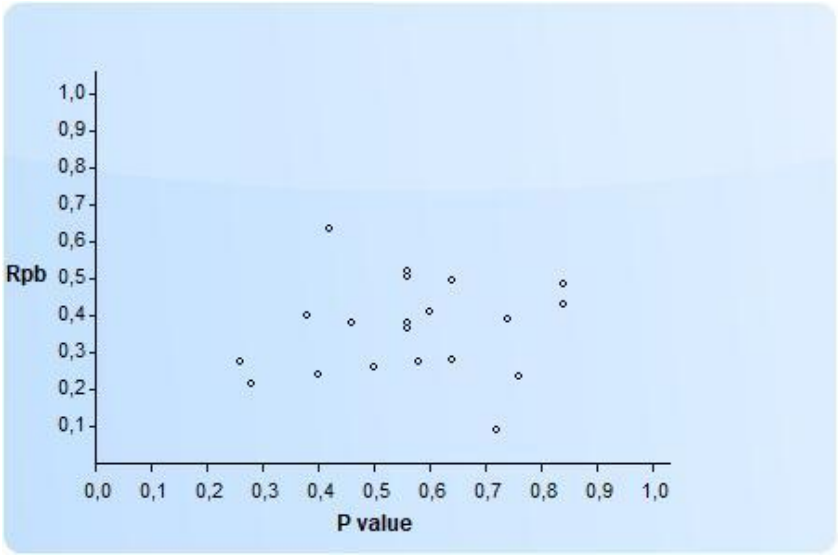


Рисунок 3.14 – Залежність трудності завдання від його можливості диференціації

Чим більше трудність завдання, тим менша міра дискримінації завдань у тесті. Аналізуючи даний графік, видно, що ця закономірність порушена. Це ще раз показує необхідність переглянути завдання експертам.

У наступному розділі звіту представлені результати аналізу кожного завдання. Для прикладу детально розглянемо завдання1.

Спочатку подано таблицю інформації про завдання (табл. 3.3).

Таблиця 3.3 – Таблиця інформації про завдання

Seq.	ID	Key	Scored	Num Options	Domain	Flags
1	Item01	3	Yes	5	1	

Завдання 1 містило 5 варіантів відповідей, третій варіант – правильний.

Далі зображено таблиця загальної статистики (табл. 3.4).

Таблиця 3.4 – Таблиця статистики завдання

N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,840	0,430	0,648	0,789

Кількість випробуваних, які відповіли на завдання 1 – 50. Рівень складності завдання – 0,840, тобто завдання легке. Значення Rpbis – 0,430, прийнятне для міри диференційованості, значення Rbis – 0,648, тобто дане завдання валідне для тесту, та коефіцієнт надійності – 0,789, що є гарним показним.

У наступній таблиці показано детальну статистику для кожного дистактора, їх правдоподібність (табл. 3.5);

Таблиця 3.5 – Статистична обробка дистракторів у завданні

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	2	0,040	-0,379	-0,861	3,000	1,414	Maroon	
2	1	0,020	-0,087	-0,253	8,000	0,000	Green	
3	42	0,840	0,430	0,648	12,214	4,064	Blue	**KEY**
4	3	0,060	-0,176	-0,350	7,667	1,528	Olive	
5	2	0,040	-0,150	-0,342	7,500	0,707	Gray	

У першому стовпчику показано варіанти відповідей, у другому – кількість випробуваних, що обрали дану відповідь, у третьому – частка випробуваних, що обрали дану відповідь. Нерівномірний розподіл обрання відповідей говорить про невдалий підбір дистракторів, тобто їх неправдоподібність [17].

Дистрактор під номером 2 має значення близьке до нуля, тому його потрібно переробити. Значення Rpbis, Rbis інших дистракторів відповідають нормам, що говорить про їх правдоподібність [19].

Таблиця 3.6 показує дані квантильного графіку – значення, для створення квантильного графіку.

Таблиця 3.6 – Дані квантильного графіку

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	2	0,222	0,000	0,000	0,000	0,000	Maroon	
2	1	0,000	0,091	0,000	0,000	0,000	Green	
3	42	0,556	0,636	1,000	1,000	1,000	Blue	**KEY**
4	3	0,111	0,182	0,000	0,000	0,000	Olive	
5	2	0,111	0,091	0,000	0,000	0,000	Gray	

Но основі значень, що представлені вище у таблиці 3.5, побудовано квантильний графік дистракторів (рис. 3.15).

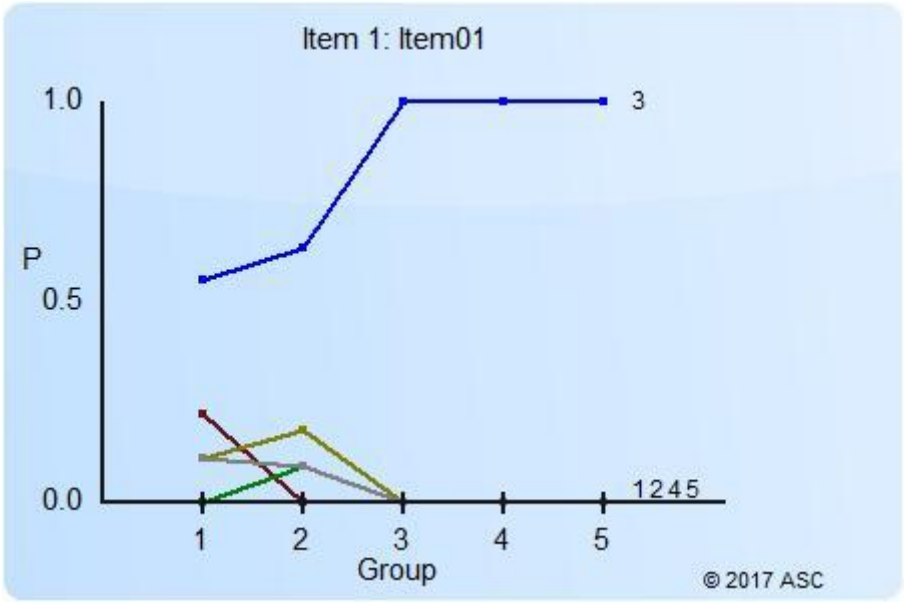


Рисунок 3.15 – Квантильний графік відповідей завдання

З графіку видно, що ті випробувані, хто відповів правильно на 40-100% (група 3 по осі абсцис) завдань обрали вірну відповідь. Для першого завдання у тесті даний результаті вважається прийнятним.

Дослідивши всі інші завдання та їх дистрактори (таблиці статистичних даних розміщено у додатку 1) можна зробити висновок, що у всіх дистракторів відсутній рівномірний розподіл обрання відповідей. Завдання 2, 5, 10, 11, 15, 16, 18, 20 мають значення Rpbis<0,3 та Rbis<0,4 що говорить про низьку спроможність диференціювання випробуваних та низьку валідність завдань тесту. Особливо завдання 5, в якому Rpbis=0,088, Rbis=0,117, не можливо вважати валідним. Його необхідно в першу чергу переглянути та корегувати.

У завданнях 5 – 7, 10, 11, 13, 15 – 18, 20 деякі неправильні відповіді мають додатні значення Rpbis та Rbis (дистрактор обрав випробуваний з високим індивідуальним балом), або взагалі нуль (жоден випробуваний не вибрав цей дистрактор). Потрібна корекція «проблемних» варіантів дистракторів завдань.

Проаналізувавши вище вказані дані обробки результатів тестування, можна стверджувати, що хоча коефіцієнт надійності достатньо високий та рівень трудності завдань у тесті відповідає нормам, але валідність та ефективність завдань тесту занизька. Для покращення останніх двох показників, потрібно виконати рекомендації та провести апробацію тесту для порівняння результатів.

3.3 Обробка результатів тестування пробного ЗНО за допомогою програми Ministep

Використовуючи початкову матриця результатів тестування з програми MS Excell, конвертовано дані та отримано загальний аналіз щодо випробуваних та завдань тесту за допомогою програми Ministep (рис. 3.16).

PERSON	50 INPUT		50 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	11.3	20.0	.43	.60	1.00	.0	1.07	.1
P.SD	4.2	.0	1.32	.22	.17	.7	.52	.7
REAL RMSE	.64	TRUE SD	1.16	SEPARATION	1.81	PERSON RELIABILITY		.77
ITEM	20 INPUT		20 MEASURED		INFIT		OUTFIT	
	TOTAL	COUNT	MEASURE	REALSE	IMNSQ	ZSTD	OMNSQ	ZSTD
MEAN	28.3	50.0	.00	.35	.99	.0	1.07	.2
P.SD	8.1	.0	.93	.03	.14	1.0	.42	1.1
REAL RMSE	.36	TRUE SD	.86	SEPARATION	2.42	ITEM RELIABILITY		.85

Рисунок 3.16 – Загальний аналіз результатів тестування

Результати щодо випробуваних співпадають з результатами, які було обраховано за допомогою програми Iteman: середнє значення, стандартне відхилення, надійність завдань тесту (значення Спірмена-Брауна). Додатково подано таку інформацію:

а) середнє значення завдань тесту = 28,3;

б) стандартне відхилення завдань тесту = 8,1;

в) рівень складності завдань у логітах = 0;

г) рівень підготовленості випробуваних у логітах = 0,43.

У стовпчиках INFIT та OUTFIT значення MNSQ, а також у полі ZSTD, що для випробуваних, що для завдань є вдалими.

Отримано детальний аналіз у вимірюваннях Раша для всіх завдань тесту, використавши команду Item: measure в меню Output Table (рис. 3.17).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
11	13	50	1.75	.37	1.14	.73	1.12	.44	.41	.49	77.6	79.9	11
10	14	50	1.61	.37	1.13	.71	1.36	1.10	.38	.49	77.6	78.4	10
13	19	50	1.00	.34	.96	-.22	1.12	.58	.49	.49	75.5	72.5	13
15	20	50	.89	.33	1.20	1.37	1.11	.52	.38	.49	57.1	71.9	15
8	21	50	.78	.33	.72	-2.18	.63	-1.87	.67	.48	83.7	71.3	8
14	23	50	.57	.32	1.01	.14	.93	-.28	.48	.48	67.3	70.6	14
20	25	50	.36	.32	1.14	1.09	1.30	1.38	.37	.47	61.2	70.1	20
3	28	50	.05	.32	1.00	.08	.87	-.52	.47	.45	59.2	69.7	3
4	28	50	.05	.32	.86	-1.18	.76	-1.13	.56	.45	75.5	69.7	4
17	28	50	.05	.32	1.01	.08	1.06	.36	.44	.45	75.5	69.7	17
19	28	50	.05	.32	.84	-1.32	.78	-1.01	.56	.45	79.6	69.7	19
16	29	50	-.05	.32	1.12	.96	1.02	.16	.39	.45	61.2	69.7	16
9	30	50	-.16	.33	.94	-.46	1.38	1.51	.45	.44	79.6	69.9	9
12	32	50	-.37	.33	.86	-1.08	.73	-1.02	.53	.43	77.6	70.7	12
18	32	50	-.37	.33	1.08	.62	1.04	.23	.38	.43	73.5	70.7	18
5	36	50	-.83	.35	1.23	1.39	1.73	1.86	.20	.40	71.4	74.8	5
6	37	50	-.95	.36	.89	-.61	.93	-.07	.44	.39	81.6	76.2	6
2	38	50	-1.08	.36	1.10	.58	2.39	2.67	.27	.38	73.5	77.8	2
1	42	50	-1.68	.42	.85	-.56	.56	-.77	.45	.33	85.7	84.4	1
7	42	50	-1.68	.42	.78	-.88	.49	-.95	.49	.33	85.7	84.4	7
MEAN	28.3	50.0	.00	.34	.99	.0	1.07	.2			74.0	73.6	
P.SD	8.1	.0	.93	.03	.14	1.0	.42	1.1			8.4	4.8	

Рисунок 3.17 – Статистичні дані завдань тесту

За результатами вимірювання складності завдань у логітах, видно, що у тесті порушено принцип диференційованість завдань. У стовпчику MODEL S.E. наведена похибка вимірювання на основі моделі Раша. Значення даного стовпчика задовільнять прийнятим нормам.

У стовпчиках INFIT та OUTFIT більшість параметрів, що характеризують відповідність даних моделі Раша розташовані у межах від 0,5 до 1,5. Завдання 2

має значення MNSQ OUTFIT=2,39, що говорить про не відповідність моделі вимірювання, тому воно не може бути використане при аналізі результатів. Завдання 5 має значення MNSQ OUTFIT=1,73, що означає угадування відповіді. Тому дане завдання потрібно переглянути та корегувати. Це ж стосується і завдання 8, яке має значення ZSTD INFIT=-2,18.

Аналізуючи стовпчик PT-MEASURE CORR, виділено завдання 2 та 5, що маю значення коефіцієнта кореляції менше за 0,3. Їх можна вважати не валідними для тесту. Необхідна заміна завдань.

Побудовано характеристичні криві усіх завдань досліджуваного тесту (рис. 3.18).

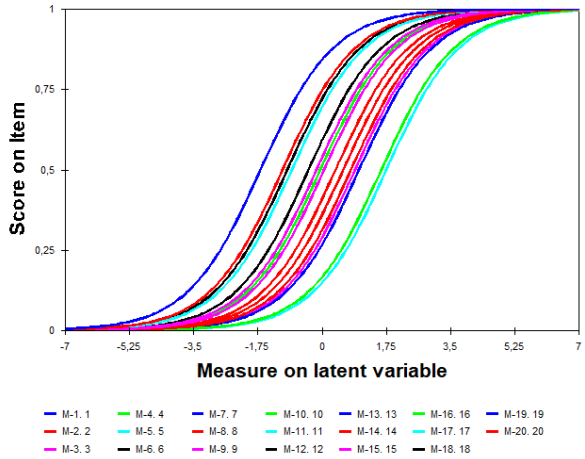


Рисунок 3.18 – Графіки характеристичних кривих завдань тесту

Отримано такий же графік характеристичних кривих завдань тесту, як і за допомогою програми MS Excell. Наочно видно, що криві не рівномірно представлені на всьому інтервалі (-7;7), накладаються одна на одну, що говорить про не коректність побудови завдань у тесті за складністю.

Для кожного завдання тесту (рис. 3.19) та тесту в цілому (рис. 3.20) побудовано графіки інформаційних функцій, які використовуються для оцінки ефективності тесту.

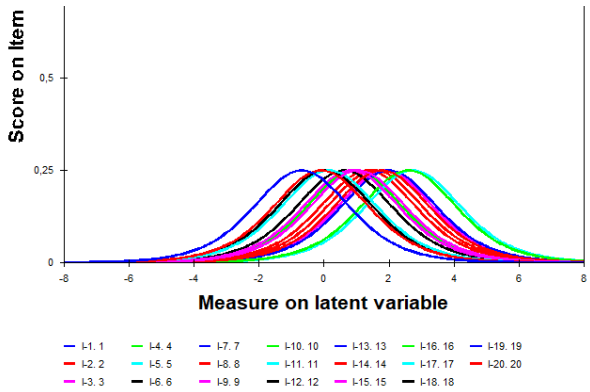


Рисунок 3.19 – Графіки інформаційних функцій кожного завдання

Більшість завдання мають рівномірно розподілену складність вздовж осі і інформаційна функція кожного завдання має один яскраво виражений екстремум, що є гарним показником для тесту. Можна виділити завдання 1 та 10, які явно порушують рівномірність розподілу. Їх потрібно ускладнити або полегшити, так щоб всі криві розміщувались рівномірно.

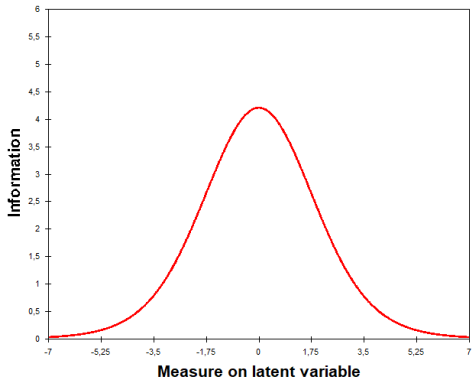


Рисунок 3.20 – Графік інформаційної функції завдань тесту

Аналізуючи останні два графіки, можна зробити висновок, незважаючи на завдання, що порушують рівномірність складності завдань, інформаційна функція тесту загалом має чітко виражений екстремум, що є показником для гарно збалансованого тесту.

Отримано звіт з характеристиками учасників тестування за допомогою командою Person: measure в меню Output Tables (рис. 3.21).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	PERSON
21	20	20	4.57	1.84	MAXIMUM MEASURE				.00	.00	100.0	100.0	студ 21
22	19	20	3.30	1.05	1.13	.43	1.65	.86	-.04	.18	95.0	95.0	студ 22
26	19	20	3.30	1.05	1.17	.47	4.07	1.74	-.27	.18	95.0	95.0	студ 26
20	18	20	2.51	.77	.92	.04	.79	.06	.32	.25	90.0	90.0	студ 20
25	18	20	2.51	.77	1.14	.42	1.75	.98	.03	.25	90.0	90.0	студ 25
3	17	20	2.01	.66	.67	-.78	.41	-.90	.64	.30	85.0	84.9	студ 3
30	16	20	1.62	.59	.82	-.46	1.02	.21	.44	.33	85.0	80.3	студ 30
6	15	20	1.29	.55	.90	-.29	.93	-.01	.43	.35	85.0	76.9	студ 6
14	15	20	1.29	.55	.98	.00	.85	-.21	.40	.35	75.0	76.9	студ 14
16	15	20	1.29	.55	.98	.00	.85	-.21	.40	.35	75.0	76.9	студ 16
18	15	20	1.29	.55	.76	-.84	.82	-.30	.55	.35	85.0	76.9	студ 18
19	15	20	1.29	.55	.76	-.84	.82	-.30	.55	.35	85.0	76.9	студ 19
43	15	20	1.29	.55	.93	-.18	.74	-.49	.46	.35	75.0	76.9	студ 43
10	14	20	1.00	.53	.81	-.79	.72	-.74	.57	.37	75.0	73.2	студ 10
23	14	20	1.00	.53	.87	-.48	.77	-.56	.51	.37	85.0	73.2	студ 23
28	14	20	1.00	.53	1.15	.70	1.19	.60	.21	.37	65.0	73.2	студ 28
29	14	20	1.00	.53	1.01	.14	.86	-.28	.39	.37	65.0	73.2	студ 29
2	13	20	.73	.51	.95	-.15	.83	-.48	.45	.38	60.0	70.6	студ 2
24	13	20	.73	.51	.96	-.15	.86	-.37	.44	.38	70.0	70.6	студ 24
27	13	20	.73	.51	.82	-.85	.72	-.91	.57	.38	80.0	70.6	студ 27
50	13	20	.73	.51	1.25	1.19	1.14	.53	.17	.38	60.0	70.6	студ 50
33	12	20	.48	.50	1.17	.95	1.16	.65	.23	.39	60.0	68.6	студ 33
36	12	20	.48	.50	.84	-.85	.76	-.93	.56	.39	70.0	68.6	студ 36
39	12	20	.48	.50	.84	-.85	.76	-.93	.56	.39	70.0	68.6	студ 39
31	11	20	.24	.49	1.05	.34	1.03	.18	.36	.40	60.0	67.0	студ 31
41	11	20	.24	.49	.89	-.61	.89	-.41	.50	.40	80.0	67.0	студ 41
4	10	20	.00	.49	.99	.00	.94	-.19	.42	.40	75.0	65.9	студ 4
15	10	20	.00	.49	.94	-.32	.86	-.59	.48	.40	55.0	65.9	студ 15
35	10	20	.00	.49	.89	-.59	.92	-.30	.49	.40	75.0	65.9	студ 35
49	10	20	.00	.49	.84	-.93	.89	-.43	.54	.40	85.0	65.9	студ 49
5	9	20	-.24	.49	.99	-.02	.95	-.12	.42	.40	80.0	67.0	студ 5
7	9	20	-.24	.49	1.15	.87	1.16	.74	.25	.40	60.0	67.0	студ 7
32	9	20	-.24	.49	.96	-.15	1.07	.38	.41	.40	80.0	67.0	студ 32
37	9	20	-.24	.49	1.22	1.25	1.15	.70	.20	.40	50.0	67.0	студ 37
42	9	20	-.24	.49	.87	-.71	.84	-.66	.53	.40	80.0	67.0	студ 42
11	8	20	-.48	.50	.82	-.95	.76	-.89	.58	.39	85.0	68.8	студ 11
12	8	20	-.48	.50	1.11	.63	1.27	1.05	.25	.39	75.0	68.8	студ 12
13	8	20	-.48	.50	1.18	.98	1.12	.52	.23	.39	55.0	68.8	студ 13
17	8	20	-.48	.50	1.21	1.09	1.15	.63	.21	.39	55.0	68.8	студ 17
34	8	20	-.48	.50	1.29	1.47	1.44	1.55	.08	.39	55.0	68.8	студ 34
46	8	20	-.48	.50	1.11	.61	1.07	.36	.30	.39	65.0	68.8	студ 46
1	7	20	-.73	.51	1.25	1.18	1.52	1.56	.09	.38	60.0	71.0	студ 1
8	7	20	-.73	.51	.80	-.93	1.11	.44	.51	.38	90.0	71.0	студ 8
44	7	20	-.73	.51	.78	-1.08	.69	-1.05	.62	.38	80.0	71.0	студ 44
38	6	20	-1.00	.53	1.00	.08	.84	-.33	.41	.37	65.0	73.4	студ 38
45	6	20	-1.00	.53	.98	.00	.90	-.16	.40	.37	75.0	73.4	студ 45
48	6	20	-1.00	.53	1.51	1.94	1.94	2.16	-.19	.37	55.0	73.4	студ 48
9	4	20	-1.62	.60	1.02	.16	.87	-.07	.34	.33	80.0	80.3	студ 9
40	4	20	-1.62	.60	1.01	.13	1.20	.52	.29	.33	80.0	80.3	студ 40
47	2	20	-2.52	.77	1.11	.37	1.21	.52	.13	.25	90.0	90.0	студ 47
MEAN	11.3	20.0	.43	.58	1.00	.0	1.07	.1			74.0	73.6	
P.SD	4.2	.0	1.32	.22	.17	.7	.52	.7			12.0	7.6	

Рисунок 3.21 – Статистичні дані учасників тестування

Виділено студента 21, який набрав максимальну кількість балів, показано його рівень підготовленості у логітах β=4,57. Для нього неможливо оцінити параметри, що характеризують відповідність даних моделі Раша.

При всіх задовільних значеннях більшості показників насторожує значення 4,07 показника MNSQ OUTFIT для учасника під номером 26. Таке велике значення вказує на угадування відповідей.

3.4 Порівняльний аналіз результатів отриманих за допомогою MS Excell, Iteman 4 та Ministep

Обробивши результати тестування (з використанням MS Excell) за алгоритмами запропонованими О. Авраменко [17] та М. Челишковою [16], можна зробити такі висновки щодо якості тесту:

а) порушено принцип диференційованості завдань;

б) гіпотеза про нормальний закон розподілу результатів тестування за критерієм Пірсона не відкинута;

в) завдання 5 (5), 14 (20), 17 (15) та 19 (10), які від'ємно корелюють з більшістю завдань, можуть мати помилки або відсутня предметна чистота змісту завдання;

г) завдання 5 (5), 17 (15), 19 (10) та 20 (11) мають низький коефіцієнт валідності. Середнє значення кореляції завдань тесту – 0,435;

д) пари завдавань 2 (7) та 4 (6), 6 (12) та 16 (8), 8 (9) та 10 (4), 10 (4) та 11 (3), у яких кореляційні значення між собою більші за 0,3, можуть перевіряти знання з однієї теми. Середній рівень трудності завдань тесту – 0,556;

е) завдання 1 (1) та 2 (7), 6 (12) та 7 (18), а також 10 (4) – 13 (19) мають однаковий рівень складності.

Дослідивши результати обробки тестування системою Iteman 4, можна виділити наступне:

а) високий коефіцієнт надійності по завданням тесту (0,8);

б) достатній рівень складності завдань тесту (0,565);

в) низький рівень валідності тесту (0,363);

г) завдання 2, 10, 11, 15, 16, 18, 20 мають значення Rpbis<0,3 та Rbis<0,4 що говорить про низьку валідність завдань тесту;

д) завдання 5, в якому Rpbis=0,088, Rbis=0,117, не можливо вважати валідним;

е) завданнях 5 – 7, 10, 11, 13, 15 – 18, 20 деякі дистрактори мають значення Rpbis та Rbis додатні або нуль, необхідна корекція «проблемних» дистракторів завдань.

Аналізуючи отримані дані щодо якості тесту за допомогою системи Ministep, можна зробити висновок:

а) оптимальний рівень складності завдань (0 у логітах);

б) середній рівень підготовленості випробуваних (0,43 у логітах);

в) відповіді на завдання 2 та 5 більшість випробуваних вгадували;

г) завдання 2 та 5 мають низький показник валідності;

д) не рівномірно побудовані завдання за складністю;

е) випробуваний 26 ймовірно вгадував відповіді на завдання.

З отриманих даних можна виділити завдання 2 та 5, які необхідно видалити або корегувати.

Кожна програма досліджує свої аспекти якості тесту, але є значення, що можна отримати одночасно в трьох або в двох системах, наведено в таблиці 3.7.

Таблиця 3.7 – Значення якості тесту

Параметри	MS Excell	Iteman	Ministep
Середнє вибіркове	11,122	11,3	11,3
Стандартне відхилення	4,126	4,273	4,2
Середня трудність завдань тесту	0,556	0,565	
Коефіцієнтів надійності по завданням (кореляція Спірмена-Брауна)		0,851	0,85

У залежності наскільки точний результат необхідний досліднику та які саме аспекти якості тесту потрібно дослідити, обирається система для аналізу.

3.4 Висновки до третього розділу

У даному розділі було проведено аналіз якості результатів тестування пробного ЗНО з математики лише тестової частини, що проводилось у навчальному закладі ЗНУ з вибіркою у 50 випробуваних одразу трьома програмами: MS Excell, Iteman 4 та Ministep. Оцінивши зручність у використанні

та витрати часу на аналіз систем, можна відзначити один вагомий недолік у всіх програмах – введення результатів власноруч, що може призвести до помилкових вимірювань.

ВИСНОВКИ

У даній роботі було розглянуто історію створення тестології, висвітлено науковців та їх внесок для розвитку науки, а також основні її поняття та критерії щодо якості тесту. Ознайомлено з основними положеннями CRT та IRT теорій, описано моделі для IRT: однопараметрична (Г. Раш), двопараметрична та трипараметрична (А. Бірнбаум). Розглянуто автоматизовані системи на основі даних теорій обробки результатів тестування та порівняно їх функціонал.

На основі викладеного в роботі матеріалу було досліджено якість пробного тесту ЗНО з математики (лише 20 тестових завдань), що проходило на базі навчального закладу ЗНУ у 2019 році за допомогою програм MS Excell, Iteman, Ministep та проведено порівняльний аналіз цих систем за результатами. Порівнювалось не лише можливості систем, а й зручність їх у використанні та інтерфейс.

Результати роботи можуть бути використанні для детального дослідження якості педагогічних тестів, покращення системи навчання, професійної підготовки фахівців.

ПЕРЕЛІК ПОСИЛАНЬ

1. Булах І. Є., Мруга М. Р. Створюємо якісний тест: навч. посіб. Київ : Майстер-клас, 2006. 160 с.

2. Булах І. Є. Історія розвитку та сучасний стан педагогічної тестології. Київ : ЦМК МОЗ України, 1994. 21 с.

3. Galton, F. Inquiries into Human Faculty and its Development. AMS Press, New York. 1883. 290 с.

4. Гильбух Ю. З. Интеллектуальное тестирование на Западе: итоги и проблемы. Советская педагогика. Москва, 1980. 136 с.

5. Гаврилюк Н.М. Моніторинг якості освіти: зарубіжний досвід // *Наукові записки Вінницького державного педагогічного університету імені Михайла Коцюбинського*. 2012. № 38. С. 350-354.

6. Цатурова И. Из истории развития тестов в СССР и за рубежом. Таганрог, 1969.

7. Ebel R. L. Educational testing: Valid? Biased? Useful? // *Phi. Delta Kappun*, 1975, October, vol. 57, N 2.

8. Resnick L. B., Melnick De. A new perspective on the use of standardized tests//*Phi. Delta Kappun*, 1981, May, N 5.

9. Мишко С.А. Становлення системи тестування у сфері освіти США в ХХ столітті // *Науковий вісник УжНУ*. 2017. №1(36). С.81-85.

10. Булах І. Є., Волосовец В. Ф., Вороненко Ю. В. Система управління якістю медичної освіти в Україні: монографія. Дніпро : АРТ-ПРЕС, 2003. 212 с.

11. Кухар Л. О., Сергієнко В. П. Конструювання тестів. Курс лекцій : навч. посіб. Луцьк, 2010. 182 с.

12. Балыхина Т. М. Словарь терминов и понятий тестологии. Москва : РУДН, 2000. 164 с.

13. Богданова М. А., Корнят В. С., Цимбала О. М. Тестові технології оцінювання знань студентів у педагогічних коледжах: метод. рекомендації. Львів : Растр-7, 2013. 44 с.

14. Вимірювання в освіті: підручник / О. В. Авраменко [та ін.]. Кіровоград : В. Ф. Лисенко, 2011. 360 с.

15. Майоров А.Н. Теория и практика создания тестов для системы образования. Москва : Интеллект, 2002. 296 с.

16. Челышкова М. Б. Теория и практика конструирования педагогических тестов: учебное пособие. Москва : Логос, 2002. 432 с.

17. Авраменко О. В., Павличенко Г. Ю., Паращук С. Д. Статистичні методи в освітніх вимірюваннях. Частина І. Класична теорія тестування : навчально-методичний посібник. Кіровоград : Лисенко В. Ф., 2012. 120 с.

18. MyTestXPro компютерное тестирование знаний: инструкция пользователя программы.

19. User Manual for Iteman 4.4. Assessment Systems Corporation. 2017. 48 p.

20. The manual for Lertap 5 URL: http://lertap5.com/Documentation/Manual.htm (дата звернення: 10.11.2019).

21. Лісова Т. В. Моделі та методи сучасної теорії тестів: навчально-методичний посібник. Ніжин: Видавець ПП Лисенко М.М., 2012. 112 с.

22. Аванесов В.С. Педагогическое измерение латентных качеств // *Педагогическая диагностика*. 2003. № 4. С. 69-77.

23. Ким В. С. Тестирование учебных достижений: монография. Уссурийск: Издательство УГПИ, 2007. 214 с.

24. Каракозов С. В., Головишников К.В. Информационно-математические модели тестирования и интерпретация результатов единого государственного экзамена // *Вестник Алтайской государственной педагогической академии*. 2003. № 3-3. С.77-99.

25. Авраменко О. В., Ковальчук Ю. О., Сергієнко В. П. та ін. Підготовка фахівців з освітніх вимірювань в Україні : [навчально-методичний комплекс] Частина 2. Ніжин : Видавець ПП Лисенко М. М., 2012. 398 с.

26. Bical. Calibrating items with the rasch model. By. Benjamin D. Wright, Ronald J. Mead and Susan R. Bell. Department of Education. University of.

27. Bock R. D., Aitkin M. Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. Psychometrika. 1981. №46. pp. 443–459.

28. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm, Journal of the Royal Statistical Society. Series B (Methodological). 1977. vol. 39 (1). pp. 1–38.

29. Ким В. С. Обработка результатов тестирования компьютерной программой RUMM-2020 // *Педагогические измерения.* 2008. №4 С.53–69.

30. Аванесов В. С. Истоки и основные понятия математической теории измерений (Item Response Theory). Статья вторая. // *Педагогические измерения*. 2007. №3. С. 3-36.

31. Смирнова Г. И. Алгоритм обработки матриц результатов тестирования с оценкой 0-1-2 и более с помощью программы RUMM-2010 // *Педагогические измерения*. 2007. №4. С. 86–90.

32. Ким В.С. Анализ тестовых заданий в модели G. RASCH // *Педагогические измерения*. 2008. №1. С.49-58.

33. Ким В.С. Анализ результатов тестирования в процессе Rasch measurement // *Педагогические измерения*. 2005. №4. С. 39-45.

34. Linacre J. M. A user’s guide to WINISTEPS, MINISTEP Rasch-Model computer program. 2006. 349 p.

ДОДАТОК А

Звіт обробки результатів тестування за допомогою програми Iteman

Table 1 presents the specifications and basic information concerning the analysis. This provides important documentation of the setup of the program for historical purposes.

Table 1: Specifications

Specification	Value	Specification	Value
Number of examinees	50	Total Items	20
Scored Items	20	Pretest Items	
Multiple Choice Items	20	Polytomous Items	0
Number of Domains	1	External scores	No
Minimum P	-0,00	Maximum P	1,00
Minimum item mean	0,00	Maximum item mean	15,00
Minimum item correlation	-0,00	Maximum item correlation	1,00
ITEMAN 3.0 Header	No	Exclude omits from option statistics	No
Number of ID columns	6	ID begins in column	1
Responses begin in column	7	Omit character	O
Not Admin character	N	Produce quantile tables	Yes
Correct for spuriousness	Yes	Produce quantile plots	Yes
Save data matrix	No	Include omit codes in matrix	N/A
Include Not Admin codes in matrix	N/A	Include scaled scores for	N/A
Scaling function	N/A	Scaled score setting 1	N/A
Scaled score new SD	N/A	Dichotomous Classification	N/A
Classify based on	N/A	Cutpoint	N/A
Low group label	Pass	High group label	Fail
Data is delimited by	Comma	Test for DIF	No
Group status is in column	N/A	Ability levels for DIF	N/A
Group 1 code	N/A	Group 2 code	N/A
Group 1 label	N/A	Group 2 label	N/A

Summary statistics

Table 2 presents the summary statistics of the test for the scored items. Definitions of these statistics are found in the Iteman manual.

Table 2: Summary statistics

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
Scored Items	20	11,300	4,273	2	20	0,565	0,363

Table 3 presents a reliability analysis of the tests. Alpha (also known as KR-20) is the most commonly used index of reliability, and is therefore used to calculate the standard error of measurement (SEM) on the raw score scale. Also presented are three configurations of split-half reliability, first as uncorrected correlations, and then as Spearman-Brown (S-B) corrected correlations. This is because an uncorrected split-half

correlation is referenced to a "test" that only contains half as many items as the full test, and therefore underestimates reliability.

Table 3: Reliability

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0,800	1,912	0,740	0,587	0,663	0,851	0,740	0,797

No items were flagged during this analysis.

Figure 1 displays the distribution of the raw scores for the scored items across all domains. Table 4 displays the frequency distribution for total score shown in Figure 1.

Figure 1: Total score for the scored items

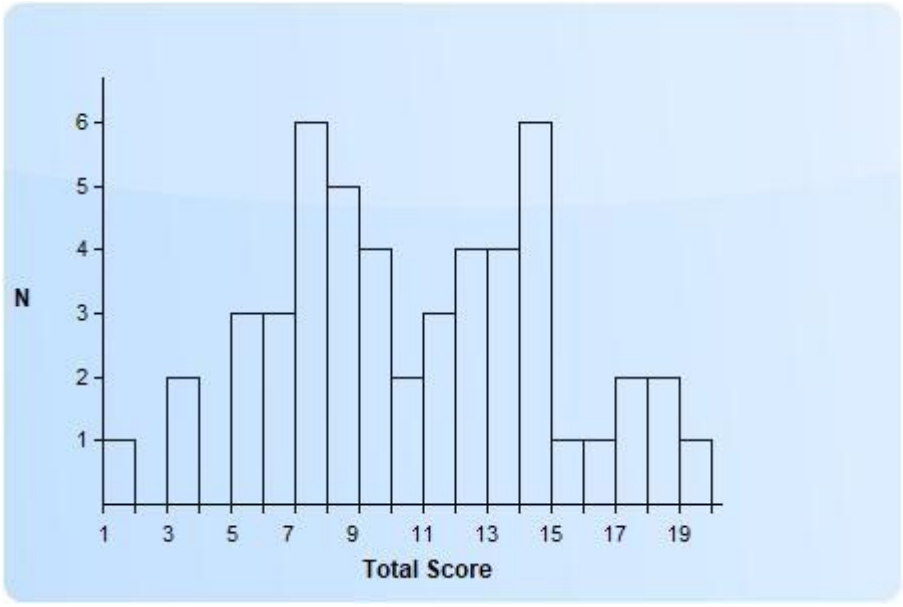


Table 4: Frequency Distribution for Total Score

Score	Frequency
2	1
3	0
4	2
5	0
6	3
7	3
8	6
9	5
10	4
11	2
12	3
13	4
14	4
15	6
16	1
17	1
18	2
19	2
20	1

Figure 2 displays the distribution of the P values for the dichotomously scored items (correct/incorrect). Table 5 displays the frequency distribution of the P values shown in Figure 2.

Figure 2: P values for the scored items

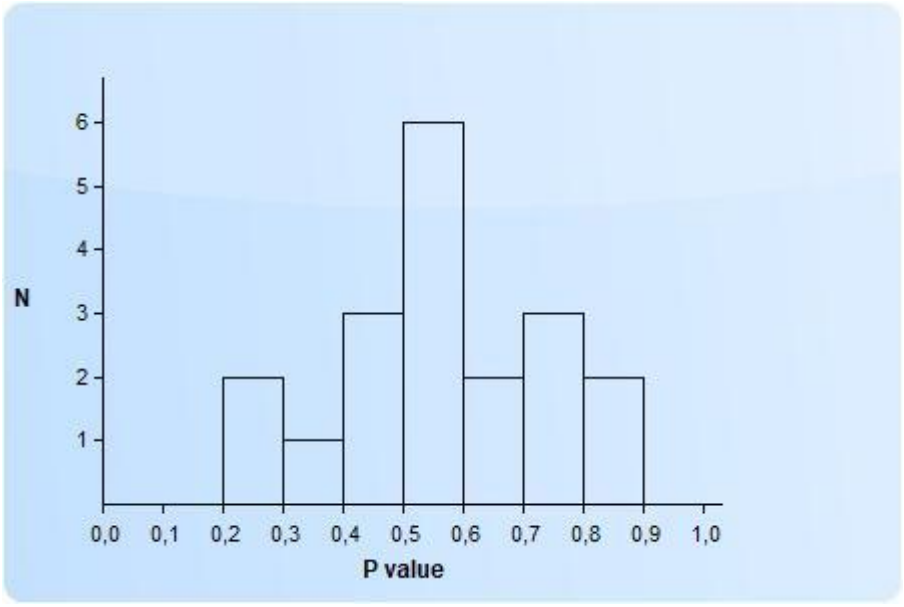


Table 5: Frequency Distribution for the P values

Score	Frequency
-0,0 to 0,1	0
0,1 to 0,2	0
0,2 to 0,3	2
0,3 to 0,4	1
0,4 to 0,5	3
0,5 to 0,6	6
0,6 to 0,7	2
0,7 to 0,8	3
0,8 to 0,9	2
0,9 to 1,0	0

Figure 3 displays the distribution of the Point-Biserial Correlations for the dichotomously scored items (correct/incorrect). Table 6 displays the frequency distribution of the Point-Biserial correlations shown in Figure 3.

Figure 3: Rpbis for the scored items

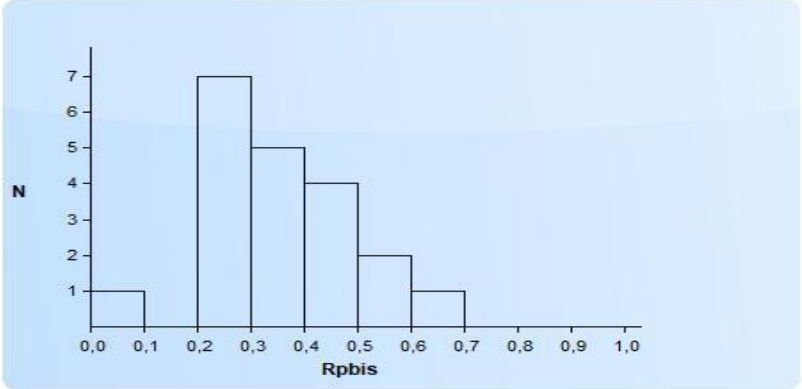


Table 6: Frequency Distribution for the Rpbis

Score	Frequency
-0,0 to 0,1	1
0,1 to 0,2	0
0,2 to 0,3	7
0,3 to 0,4	5
0,4 to 0,5	4
0,5 to 0,6	2
0,6 to 0,7	1
0,7 to 0,8	0
0,8 to 0,9	0
0,9 to 1,0	0

Figure 4 displays the scatterplot of P (difficulty) by Rpbis (discrimination) for the dichotomously scored items (correct/incorrect).

Figure 4: P by Rpbis

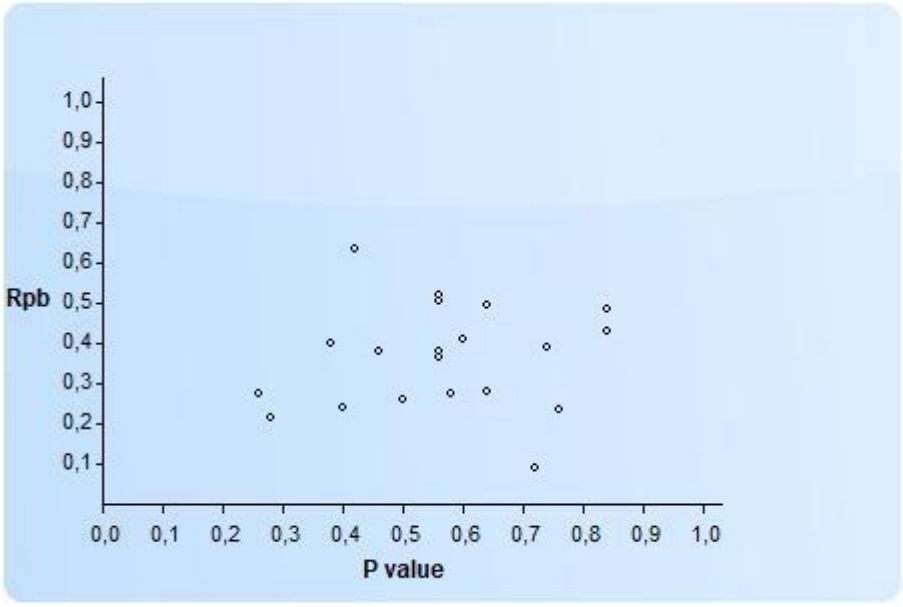
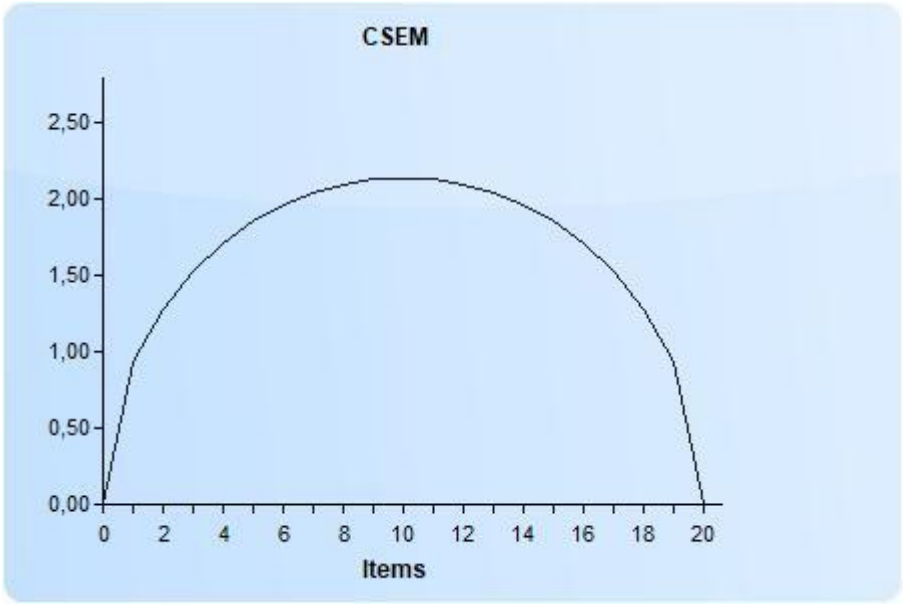


Figure 5 displays a graph of the Conditional Standard Error of Measurement (CSEM) Formula IV.

Figure 5: CSEM



Item-by-item results

The following section presents the item-by-item results of the analysis. Each item has several tables and a figure. The figure, called a quantile plot, shows the proportion of examinees selecting each option, for consecutive segments of the examinees as ranked by score. The key thing to evaluate in this figure is that the line for the correct answer has a positive slope (goes up from left to right), which means that examinees with higher scores tend to answer correctly more often. Conversely, the lines for the incorrect options, called distractors, should have a negative slope. Note, however, that the use of a small number of groups (e.g., 3 or fewer) oversimplifies the graph, so that items which are very difficult or very easy (that is, discriminating in only the top or bottom 20% of examinees) might appear to have poor quantile plots and classical statistics. For such items, item response theory presents significant advantages in analysis.

There are four tables presented for each item.

1. Item information table: records the information supplied by the control file (or Iteman 3 header) for this item.

2. Item statistics table: overall item statistics.

3. Option statistics: detailed statistics for each item, which helps diagnose issues in items with poor statistics.

4. Quantile plot data: the values used to create the quantile plot.

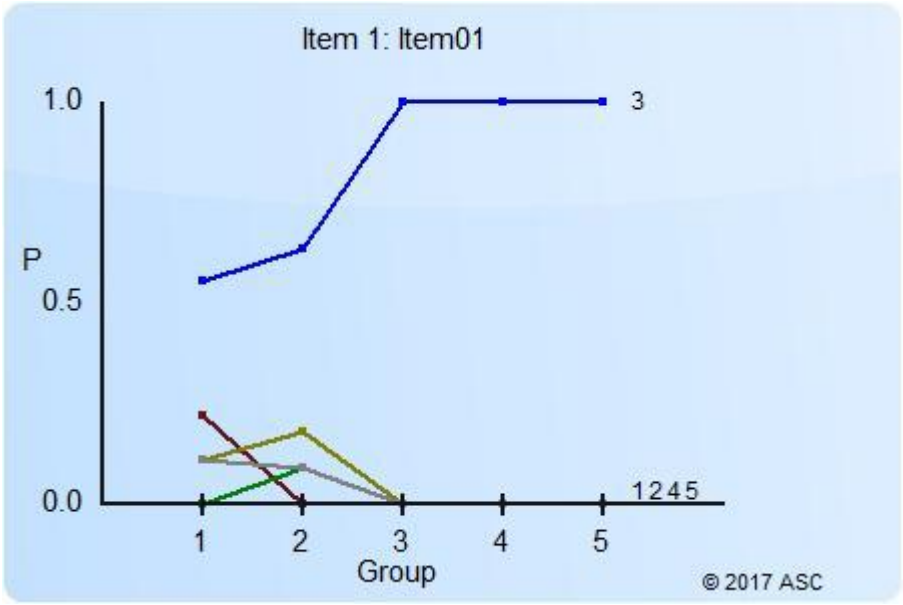
The item statistics table presents overall item statistics in the first row of numbers. The two most important item-level statistics for dichotomously scored (correct/incorrect) items are the P value and the point-biserial correlation, which represent the difficulty and discrimination of the item, respectively. For polytomously scored (rating scale or partial credit) items, the difficulty is represented by the mean (average) item score, while the discrimination is represented by a Pearson r correlation.

The P value is the proportion of examinees that answered an item in the keyed direction. P ranges from 0 to 1. A high value (0.95) means that an item is easy, a low value (0.25) means that the item is difficult. The point-biserial correlation (Rpbis) is a measure of the discriminating, or differentiating, power of the item. Rpbis ranges from -1 to 1. A negative Rpbis is indicative of a bad item as lower scoring examinees are more likely than higher scoring examinees to respond in the keyed direction.

For rating scale or partial credit items, the mean item score ranges from the minimum to the maximum of the scale. For example, if the item has a rating scale of 1 to 5, the possible range for the mean is 1 to 5. The Pearson r is similar to the Rpbis in that it ranges from -1 to 1, with a positive r indicating that the item correlates well with total score.

The option statistics table presents statistics for each individual option (alternative). The key thing to examine in this portion of the table is that no distractors have a higher Rpbis than the correct answer. That indicates that higher scoring examinees are selecting the incorrect answer, which therefore might be arguably correct.

The quantile plot data table simply presents the values calculated to create the quantile plot. Because it contains the same information, the quantile plot itself presents a useful picture of the item's performance, but this table can be used to examine that performance in detail to help diagnose possible issues.



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
1	Item01	3	Yes	5	1	

Item statistics

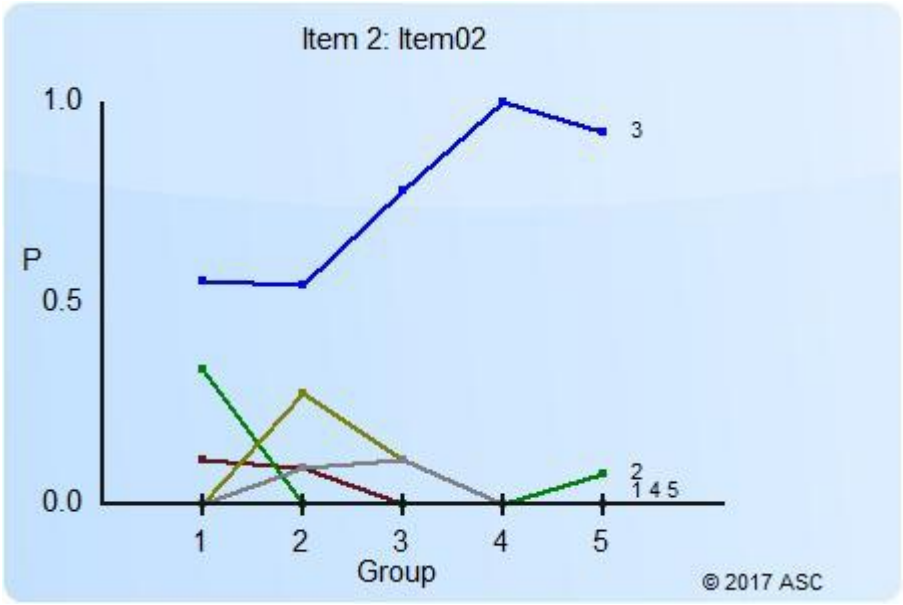
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,840	0,430	0,648	0,789

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	2	0,040	-0,379	-0,861	3,000	1,414	Maroon	
2	1	0,020	-0,087	-0,253	8,000	0,000	Green	
3	42	0,840	0,430	0,648	12,214	4,064	Blue	**KEY**
4	3	0,060	-0,176	-0,350	7,667	1,528	Olive	
5	2	0,040	-0,150	-0,342	7,500	0,707	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	2	0,222	0,000	0,000	0,000	0,000	Maroon	
2	1	0,000	0,091	0,000	0,000	0,000	Green	
3	42	0,556	0,636	1,000	1,000	1,000	Blue	**KEY**
4	3	0,111	0,182	0,000	0,000	0,000	Olive	
5	2	0,111	0,091	0,000	0,000	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
2	Item02	3	Yes	5	1	

Item statistics

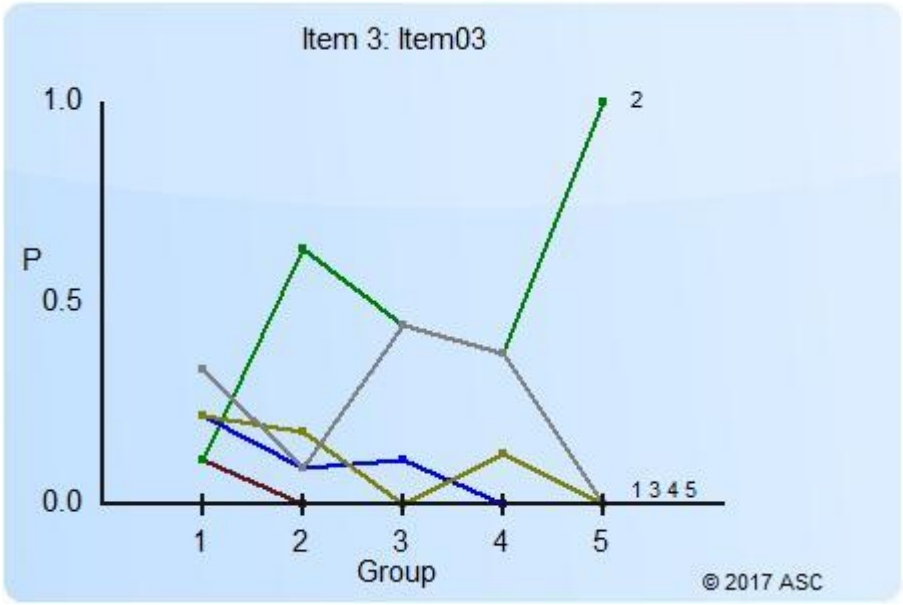
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,760	0,236	0,324	0,798

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	2	0,040	-0,228	-0,518	6,000	2,828	Maroon	
2	4	0,080	-0,075	-0,138	9,500	6,351	Green	
3	38	0,760	0,236	0,324	12,079	4,296	Blue	**KEY**
4	4	0,080	-0,094	-0,171	9,250	0,500	Olive	
5	2	0,040	-0,052	-0,119	9,500	2,121	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	2	0,111	0,091	0,000	0,000	0,000	Maroon	
2	4	0,333	0,000	0,000	0,000	0,077	Green	
3	38	0,556	0,545	0,778	1,000	0,923	Blue	**KEY**
4	4	0,000	0,273	0,111	0,000	0,000	Olive	
5	2	0,000	0,091	0,111	0,000	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
3	Item03	2	Yes	5	1	

Item statistics

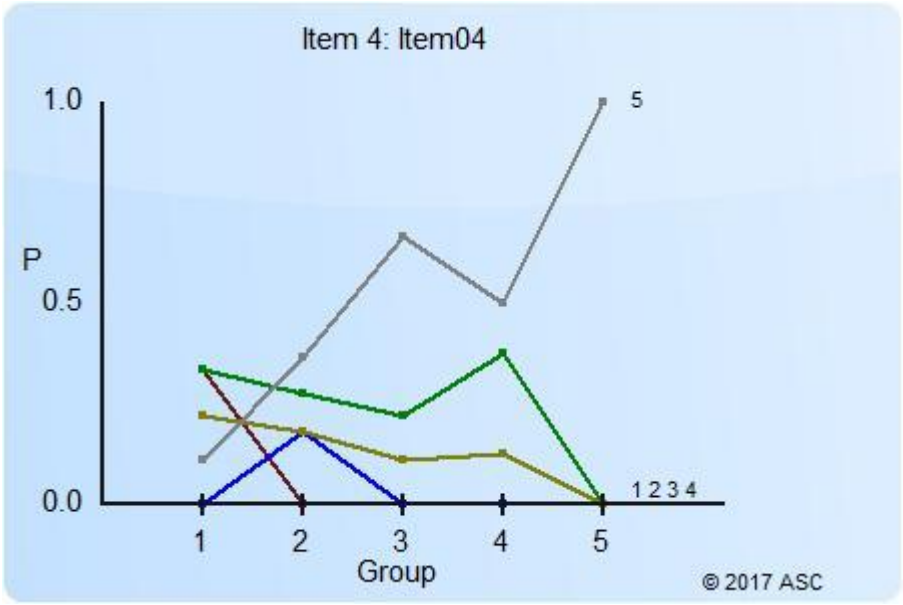
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,560	0,378	0,475	0,790

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	2	0,040	-0,115	-0,261	8,500	6,364	Maroon	
2	28	0,560	0,378	0,475	13,071	4,180	Green	**KEY**
3	4	0,080	-0,166	-0,303	8,500	1,915	Blue	
4	5	0,100	-0,196	-0,335	8,400	2,881	Olive	
5	11	0,220	-0,147	-0,206	9,636	3,982	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	2	0,111	0,000	0,000	0,125	0,000	Maroon	
2	28	0,111	0,636	0,444	0,375	1,000	Green	**KEY**
3	4	0,222	0,091	0,111	0,000	0,000	Blue	
4	5	0,222	0,182	0,000	0,125	0,000	Olive	
5	11	0,333	0,091	0,444	0,375	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
4	Item04	5	Yes	5	1	

Item statistics

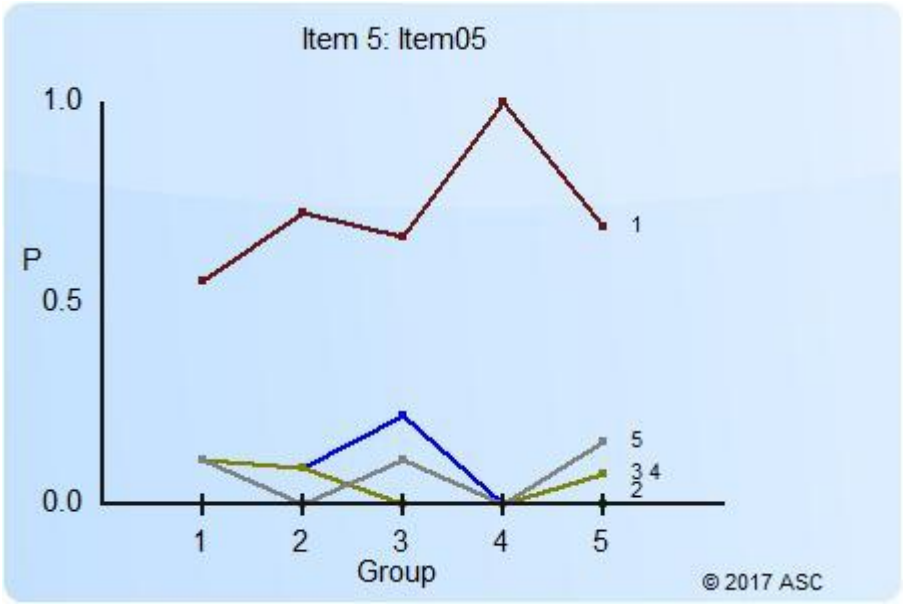
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,560	0,507	0,638	0,782

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	3	0,060	-0,392	-0,781	4,667	1,155	Maroon	
2	11	0,220	-0,187	-0,261	9,364	3,641	Green	
3	2	0,040	-0,143	-0,325	8,000	0,000	Blue	
4	6	0,120	-0,164	-0,266	9,000	2,449	Olive	
5	28	0,560	0,507	0,638	13,500	3,825	Gray	**KEY**
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	3	0,333	0,000	0,000	0,000	0,000	Maroon	
2	11	0,333	0,273	0,222	0,375	0,000	Green	
3	2	0,000	0,182	0,000	0,000	0,000	Blue	
4	6	0,222	0,182	0,111	0,125	0,000	Olive	
5	28	0,111	0,364	0,667	0,500	1,000	Gray	**KEY**



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
5	Item05	1	Yes	5	1	

Item statistics

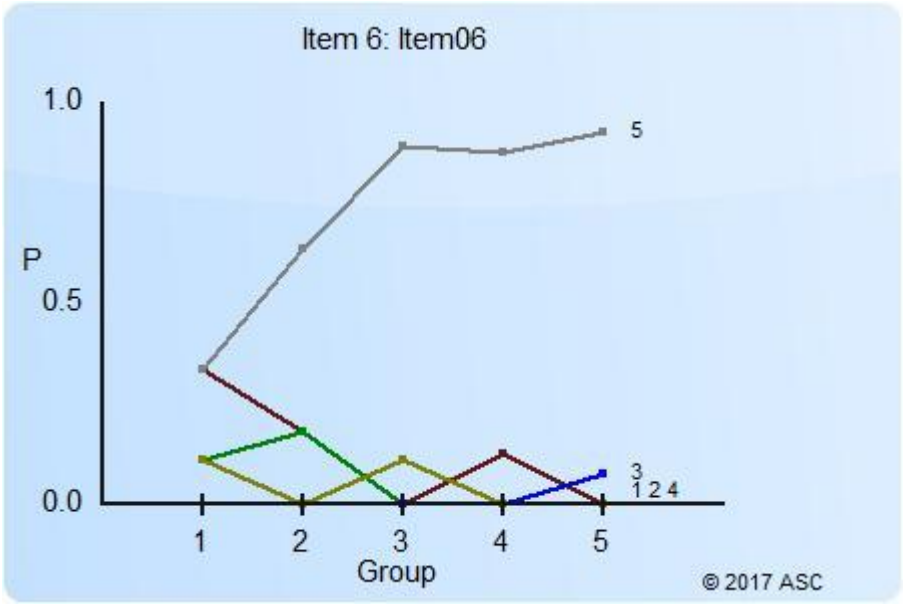
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,720	0,088	0,117	0,806

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	36	0,720	0,088	0,117	11,806	4,221	Maroon	**KEY**
2	2	0,040	-0,202	-0,459	6,500	3,536	Green	
3	5	0,100	0,018	0,030	10,800	4,207	Blue	
4	3	0,060	-0,056	-0,112	9,667	4,726	Olive	
5	4	0,080	0,030	0,055	11,000	6,164	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	36	0,556	0,727	0,667	1,000	0,692	Maroon	**KEY**
2	2	0,111	0,091	0,000	0,000	0,000	Green	
3	5	0,111	0,091	0,222	0,000	0,077	Blue	
4	3	0,111	0,091	0,000	0,000	0,077	Olive	
5	4	0,111	0,000	0,111	0,000	0,154	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
6	Item06	5	Yes	5	1	

Item statistics

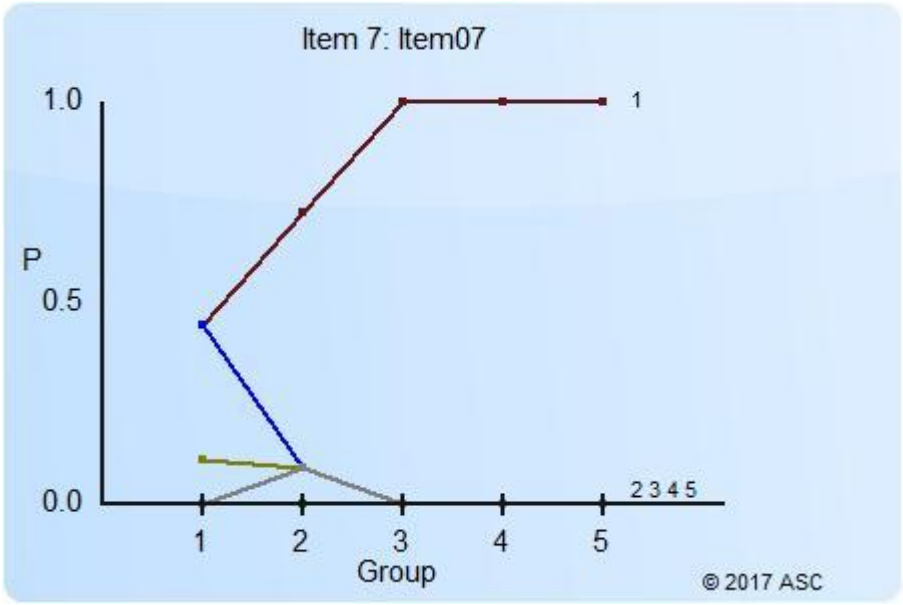
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,740	0,391	0,528	0,790

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	6	0,120	-0,282	-0,459	7,500	3,782	Maroon	
2	3	0,060	-0,183	-0,364	7,667	0,577	Green	
3	2	0,040	0,022	0,051	11,000	7,071	Blue	
4	2	0,040	-0,207	-0,471	6,500	6,364	Olive	
5	37	0,740	0,391	0,528	12,486	3,888	Gray	**KEY**
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	6	0,333	0,182	0,000	0,125	0,000	Maroon	
2	3	0,111	0,182	0,000	0,000	0,000	Green	
3	2	0,111	0,000	0,000	0,000	0,077	Blue	
4	2	0,111	0,000	0,111	0,000	0,000	Olive	
5	37	0,333	0,636	0,889	0,875	0,923	Gray	**KEY**



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
7	Item07	1	Yes	5	1	

Item statistics

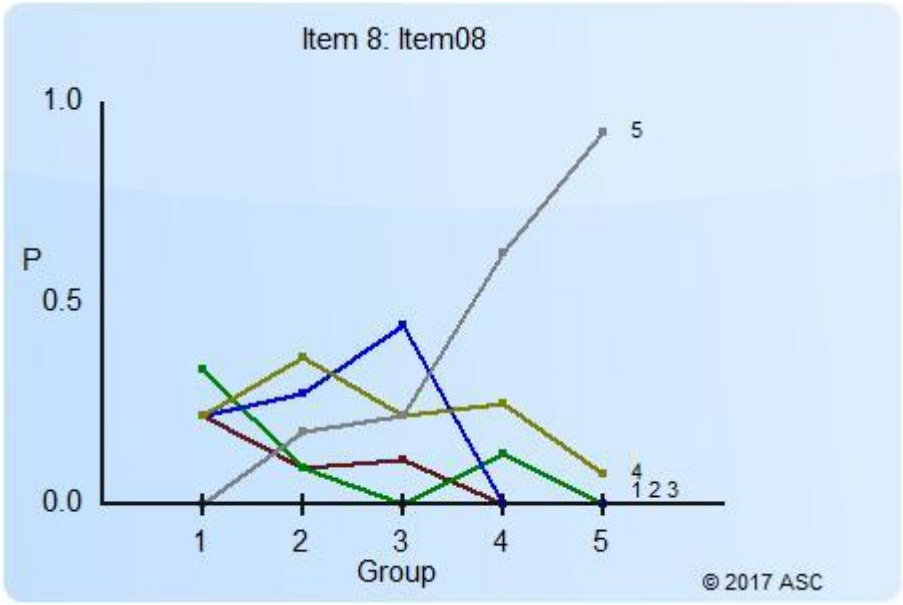
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,840	0,487	0,733	0,786

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	42	0,840	0,487	0,733	12,310	3,942	Maroon	**KEY**
2	0	0.000	--	--	--	--	Green	
3	5	0,100	-0,438	-0,749	5,200	2,280	Blue	
4	2	0,040	-0,177	-0,401	7,000	1,414	Olive	
5	1	0,020	-0,088	-0,254	8,000	0,000	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	42	0,444	0,727	1,000	1,000	1,000	Maroon	**KEY**
2	0	0,000	0,000	0,000	0,000	0,000	Green	
3	5	0,444	0,091	0,000	0,000	0,000	Blue	
4	2	0,111	0,091	0,000	0,000	0,000	Olive	
5	1	0,000	0,091	0,000	0,000	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
8	Item08	5	Yes	5	1	

Item statistics

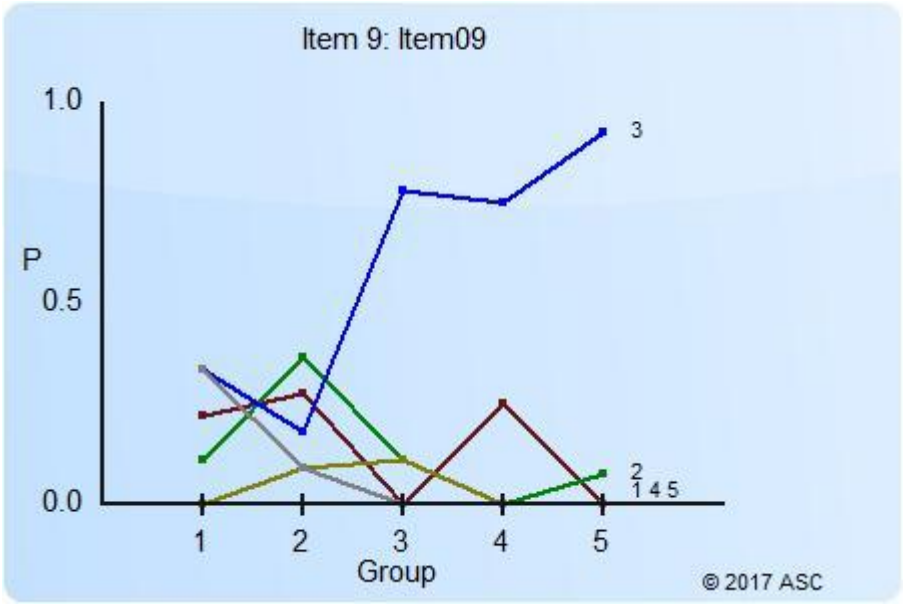
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,420	0,635	0,802	0,774

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	4	0,080	-0,258	-0,471	7,500	2,646	Maroon	
2	5	0,100	-0,249	-0,425	8,000	2,915	Green	
3	9	0,180	-0,255	-0,373	8,778	2,991	Blue	
4	11	0,220	-0,171	-0,239	9,636	3,472	Olive	
5	21	0,420	0,635	0,802	14,762	3,292	Gray	**KEY**
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	4	0,222	0,091	0,111	0,000	0,000	Maroon	
2	5	0,333	0,091	0,000	0,125	0,000	Green	
3	9	0,222	0,273	0,444	0,000	0,000	Blue	
4	11	0,222	0,364	0,222	0,250	0,077	Olive	
5	21	0,000	0,182	0,222	0,625	0,923	Gray	**KEY**



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
9	Item09	3	Yes	5	1	

Item statistics

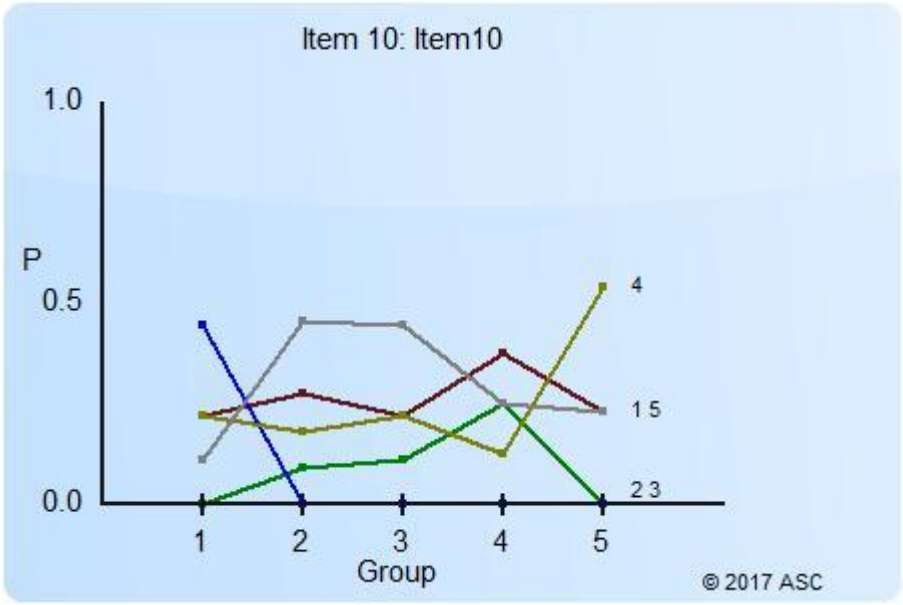
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,600	0,412	0,522	0,788

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	7	0,140	-0,173	-0,270	9,000	3,464	Maroon	
2	7	0,140	-0,115	-0,179	9,571	4,577	Green	
3	30	0,600	0,412	0,522	13,033	3,873	Blue	**KEY**
4	2	0,040	-0,036	-0,082	10,000	1,414	Olive	
5	4	0,080	-0,349	-0,638	6,000	2,944	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	7	0,222	0,273	0,000	0,250	0,000	Maroon	
2	7	0,111	0,364	0,111	0,000	0,077	Green	
3	30	0,333	0,182	0,778	0,750	0,923	Blue	**KEY**
4	2	0,000	0,091	0,111	0,000	0,000	Olive	
5	4	0,333	0,091	0,000	0,000	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
10	Item10	4	Yes	5	1	

Item statistics

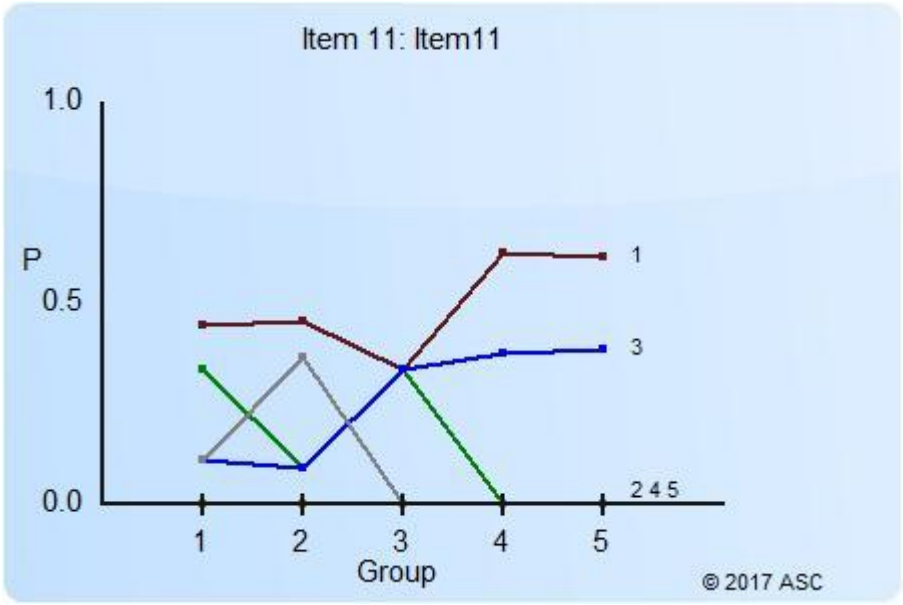
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,280	0,216	0,288	0,799

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	13	0,260	-0,003	-0,004	11,000	4,203	Maroon	
2	4	0,080	0,053	0,097	11,750	2,217	Green	
3	4	0,080	-0,400	-0,730	5,500	2,380	Blue	
4	14	0,280	0,216	0,288	13,429	5,080	Olive	**KEY**
5	15	0,300	-0,003	-0,004	11,000	3,024	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	13	0,222	0,273	0,222	0,375	0,231	Maroon	
2	4	0,000	0,091	0,111	0,250	0,000	Green	
3	4	0,444	0,000	0,000	0,000	0,000	Blue	
4	14	0,222	0,182	0,222	0,125	0,538	Olive	**KEY**
5	15	0,111	0,455	0,444	0,250	0,231	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
11	Item11	3	Yes	5	1	

Item statistics

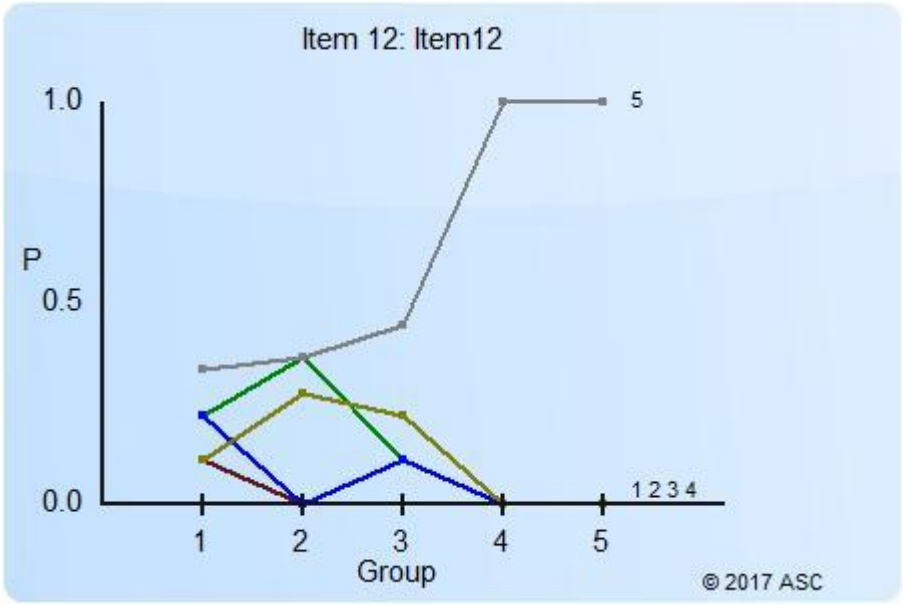
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,260	0,276	0,373	0,796

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	25	0,500	0,109	0,136	11,480	4,144	Maroon	
2	7	0,140	-0,260	-0,406	8,429	3,047	Green	
3	13	0,260	0,276	0,373	13,923	4,281	Blue	**KEY**
4	0	0.000	--	--	--	--	Olive	
5	5	0,100	-0,283	-0,484	7,600	0,894	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	25	0,444	0,455	0,333	0,625	0,615	Maroon	
2	7	0,333	0,091	0,333	0,000	0,000	Green	
3	13	0,111	0,091	0,333	0,375	0,385	Blue	**KEY**
4	0	0,000	0,000	0,000	0,000	0,000	Olive	
5	5	0,111	0,364	0,000	0,000	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
12	Item12	5	Yes	5	1	

Item statistics

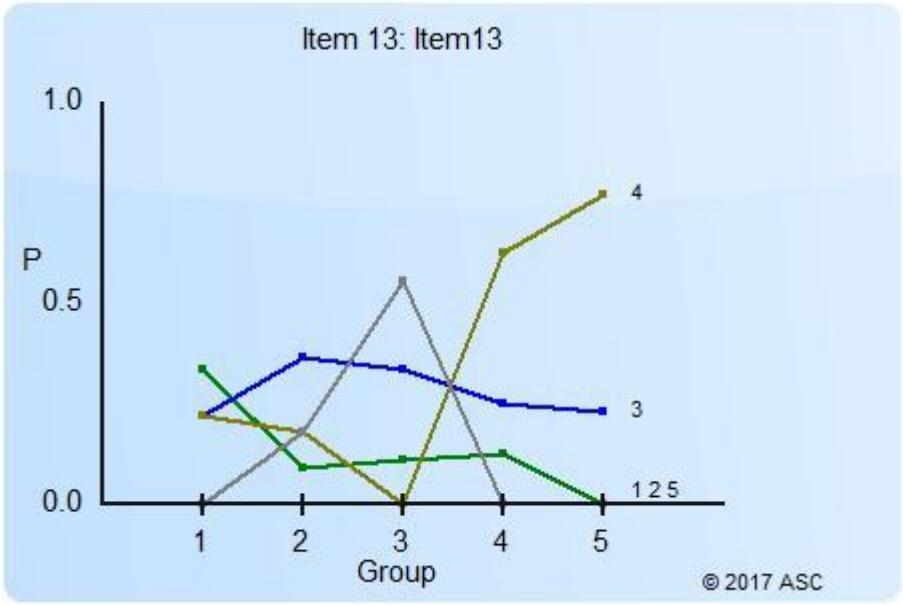
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,640	0,497	0,637	0,783

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	2	0,040	-0,164	-0,373	7,500	4,950	Maroon	
2	7	0,140	-0,331	-0,517	7,429	2,992	Green	
3	3	0,060	-0,128	-0,255	8,667	2,887	Blue	
4	6	0,120	-0,187	-0,304	8,667	1,506	Olive	
5	32	0,640	0,497	0,637	13,125	4,086	Gray	**KEY**
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	2	0,111	0,000	0,111	0,000	0,000	Maroon	
2	7	0,222	0,364	0,111	0,000	0,000	Green	
3	3	0,222	0,000	0,111	0,000	0,000	Blue	
4	6	0,111	0,273	0,222	0,000	0,000	Olive	
5	32	0,333	0,364	0,444	1,000	1,000	Gray	**KEY**



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
13	Item13	4	Yes	5	1	

Item statistics

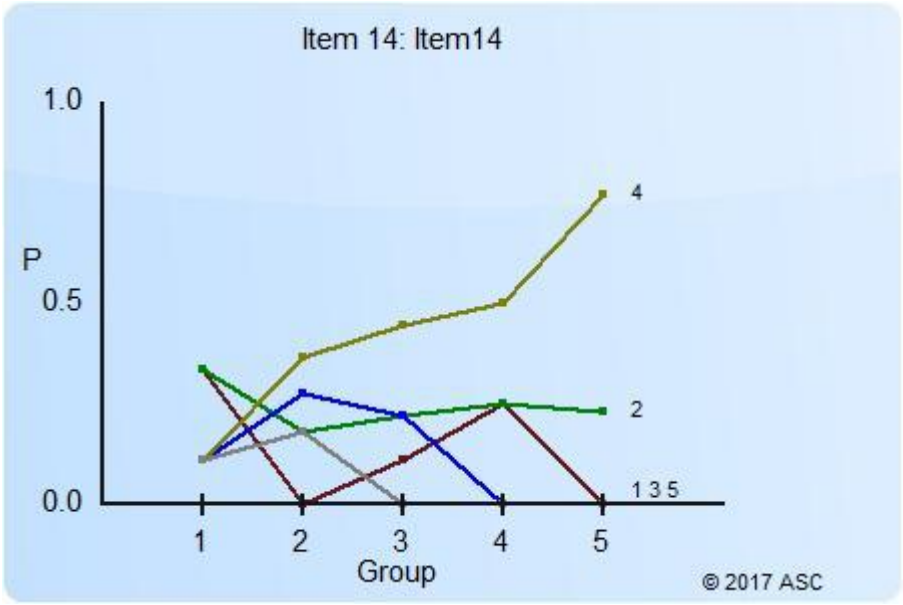
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,380	0,399	0,509	0,789

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	4	0,080	-0,328	-0,599	6,500	1,915	Maroon	
2	6	0,120	-0,256	-0,416	8,167	3,971	Green	
3	14	0,280	0,024	0,032	11,071	3,496	Blue	
4	19	0,380	0,399	0,509	13,947	4,434	Olive	**KEY**
5	7	0,140	-0,093	-0,146	10,000	1,528	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	4	0,222	0,182	0,000	0,000	0,000	Maroon	
2	6	0,333	0,091	0,111	0,125	0,000	Green	
3	14	0,222	0,364	0,333	0,250	0,231	Blue	
4	19	0,222	0,182	0,000	0,625	0,769	Olive	**KEY**
5	7	0,000	0,182	0,556	0,000	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
14	Item14	4	Yes	5	1	

Item statistics

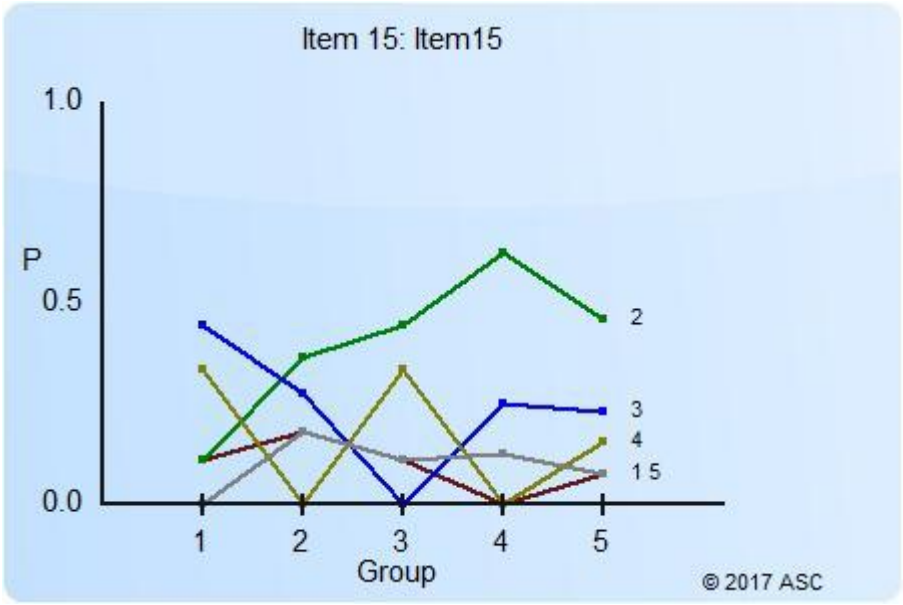
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,460	0,380	0,477	0,790

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	6	0,120	-0,155	-0,252	9,167	4,215	Maroon	
2	12	0,240	-0,036	-0,050	10,583	3,919	Green	
3	6	0,120	-0,171	-0,278	9,000	1,414	Blue	
4	23	0,460	0,380	0,477	13,478	4,116	Olive	**KEY**
5	3	0,060	-0,286	-0,571	6,333	3,786	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	6	0,333	0,000	0,111	0,250	0,000	Maroon	
2	12	0,333	0,182	0,222	0,250	0,231	Green	
3	6	0,111	0,273	0,222	0,000	0,000	Blue	
4	23	0,111	0,364	0,444	0,500	0,769	Olive	**KEY**
5	3	0,111	0,182	0,000	0,000	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
15	Item15	2	Yes	5	1	

Item statistics

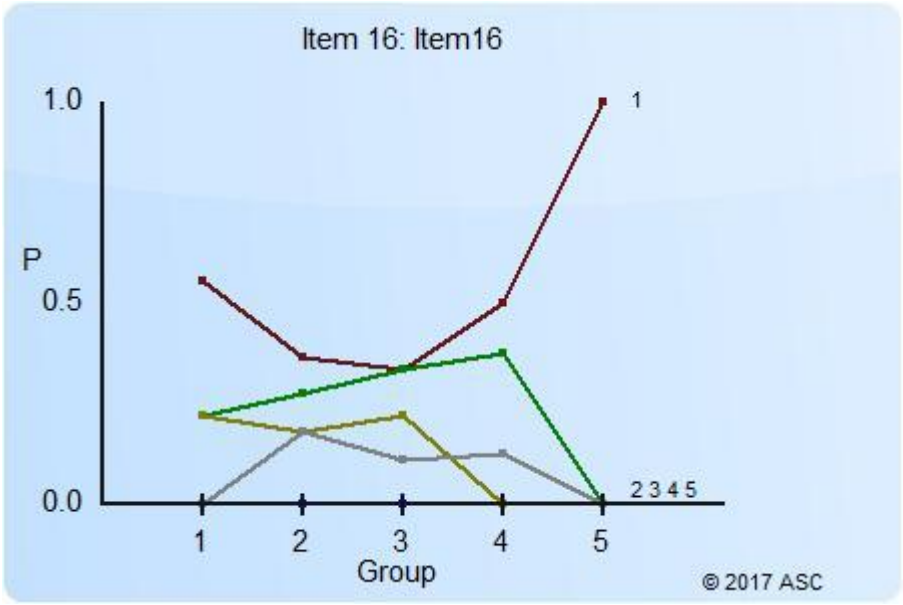
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,400	0,242	0,307	0,799

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	5	0,100	-0,140	-0,240	9,200	3,962	Maroon	
2	20	0,400	0,242	0,307	13,100	4,166	Green	**KEY**
3	12	0,240	-0,148	-0,204	9,833	4,783	Blue	
4	8	0,160	-0,084	-0,126	10,125	3,682	Olive	
5	5	0,100	0,058	0,099	11,600	3,578	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	5	0,111	0,182	0,111	0,000	0,077	Maroon	
2	20	0,111	0,364	0,444	0,625	0,462	Green	**KEY**
3	12	0,444	0,273	0,000	0,250	0,231	Blue	
4	8	0,333	0,000	0,333	0,000	0,154	Olive	
5	5	0,000	0,182	0,111	0,125	0,077	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
16	Item16	1	Yes	5	1	

Item statistics

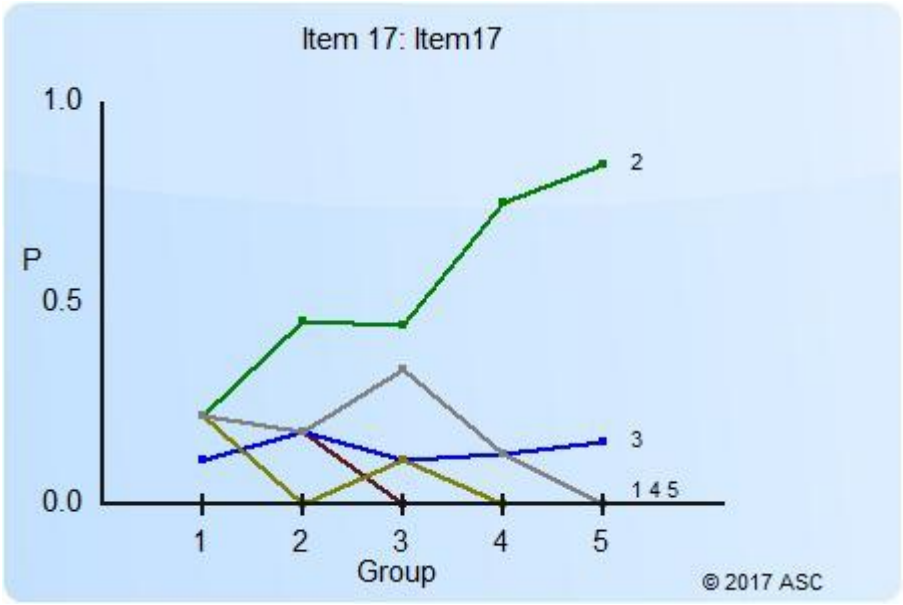
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,580	0,273	0,344	0,797

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	29	0,580	0,273	0,344	12,655	4,643	Maroon	**KEY**
2	11	0,220	-0,083	-0,116	10,091	3,113	Green	
3	0	0.000	--	--	--	--	Blue	
4	6	0,120	-0,280	-0,455	7,667	3,141	Olive	
5	4	0,080	-0,034	-0,063	10,250	2,630	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	29	0,556	0,364	0,333	0,500	1,000	Maroon	**KEY**
2	11	0,222	0,273	0,333	0,375	0,000	Green	
3	0	0,000	0,000	0,000	0,000	0,000	Blue	
4	6	0,222	0,182	0,222	0,000	0,000	Olive	
5	4	0,000	0,182	0,111	0,125	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
17	Item17	2	Yes	5	1	

Item statistics

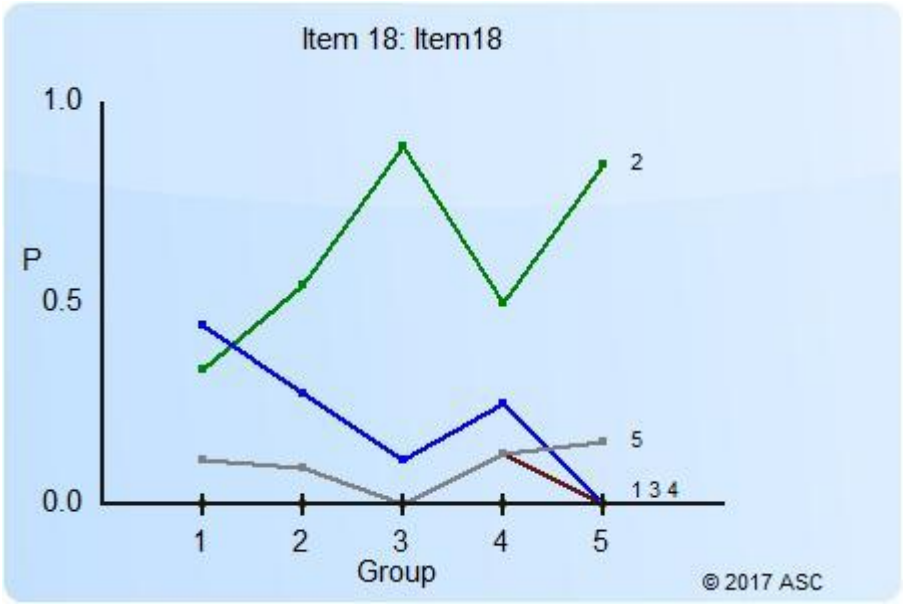
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,560	0,367	0,462	0,791

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	4	0,080	-0,332	-0,607	6,250	2,630	Maroon	
2	28	0,560	0,367	0,462	13,036	4,069	Green	**KEY**
3	7	0,140	0,070	0,109	11,429	4,036	Blue	
4	3	0,060	-0,301	-0,599	6,000	4,000	Olive	
5	8	0,160	-0,122	-0,184	9,625	2,134	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	4	0,222	0,182	0,000	0,000	0,000	Maroon	
2	28	0,222	0,455	0,444	0,750	0,846	Green	**KEY**
3	7	0,111	0,182	0,111	0,125	0,154	Blue	
4	3	0,222	0,000	0,111	0,000	0,000	Olive	
5	8	0,222	0,182	0,333	0,125	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
18	Item18	2	Yes	5	1	

Item statistics

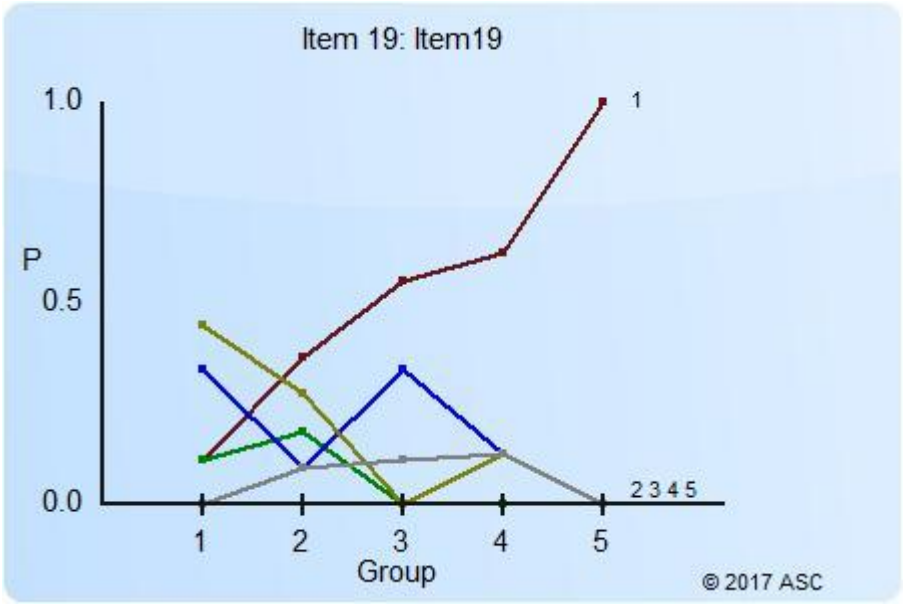
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,640	0,278	0,356	0,796

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	3	0,060	-0,062	-0,124	9,667	3,786	Maroon	
2	32	0,640	0,278	0,356	12,500	4,111	Green	**KEY**
3	10	0,200	-0,367	-0,525	7,700	3,743	Blue	
4	0	0.000	--	--	--	--	Olive	
5	5	0,100	0,094	0,161	11,800	3,962	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	3	0,111	0,091	0,000	0,125	0,000	Maroon	
2	32	0,333	0,545	0,889	0,500	0,846	Green	**KEY**
3	10	0,444	0,273	0,111	0,250	0,000	Blue	
4	0	0,000	0,000	0,000	0,000	0,000	Olive	
5	5	0,111	0,091	0,000	0,125	0,154	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
19	Item19	1	Yes	5	1	

Item statistics

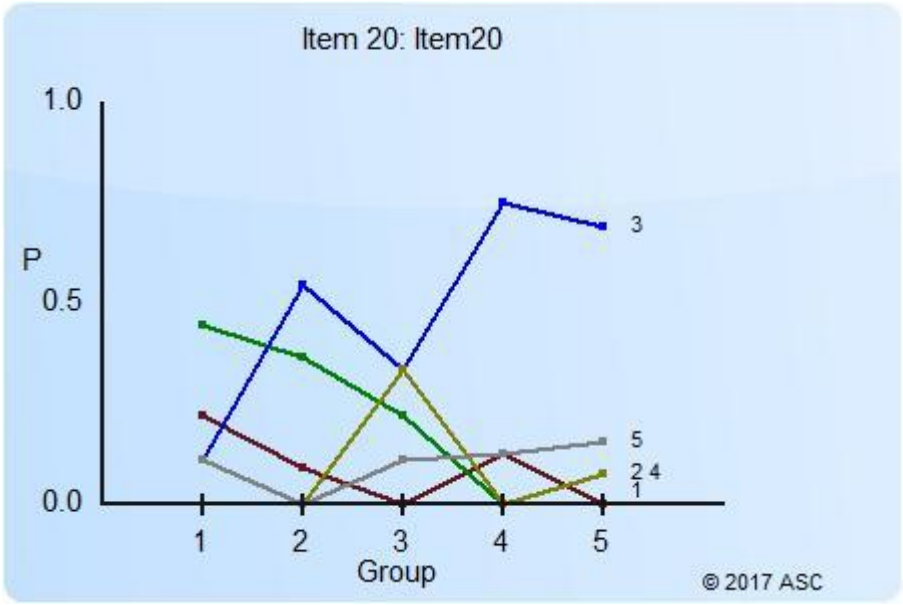
N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,560	0,518	0,651	0,782

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	28	0,560	0,518	0,651	13,536	3,967	Maroon	**KEY**
2	3	0,060	-0,198	-0,395	7,667	0,577	Green	
3	8	0,160	-0,222	-0,334	8,750	3,732	Blue	
4	8	0,160	-0,333	-0,502	7,750	2,712	Olive	
5	3	0,060	-0,026	-0,052	10,333	2,517	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	28	0,111	0,364	0,556	0,625	1,000	Maroon	**KEY**
2	3	0,111	0,182	0,000	0,000	0,000	Green	
3	8	0,333	0,091	0,333	0,125	0,000	Blue	
4	8	0,444	0,273	0,000	0,125	0,000	Olive	
5	3	0,000	0,091	0,111	0,125	0,000	Gray	



Item information

Seq.	ID	Key	Scored	Num Options	Domain	Flags
20	Item20	3	Yes	5	1	

Item statistics

N	P	Total Rpbis	Total Rbis	Alpha w/o
50	0,500	0,258	0,323	0,798

Option statistics

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
1	4	0,080	-0,205	-0,374	8,000	4,320	Maroon	
2	11	0,220	-0,273	-0,382	8,727	2,901	Green	
3	25	0,500	0,258	0,323	12,840	4,548	Blue	**KEY**
4	5	0,100	0,033	0,057	11,200	3,271	Olive	
5	5	0,100	0,099	0,170	12,000	3,937	Gray	
Omit	0							
Not Admin	0							

Quantile plot data

Option	N	0-20%	20-40%	40-60%	60-80%	80-100%	Color	
1	4	0,222	0,091	0,000	0,125	0,000	Maroon	
2	11	0,444	0,364	0,222	0,000	0,077	Green	
3	25	0,111	0,545	0,333	0,750	0,692	Blue	**KEY**
4	5	0,111	0,000	0,333	0,000	0,077	Olive	
5	5	0,111	0,000	0,111	0,125	0,154	Gray	