6. Жигалов В.Т., Шимановська Л.М. Основи менеджменту і управлінської діяльності. — К.: Вища шк., 1994. — 223 с.
7. Ступницький О. Інформаційні технології та корпоративне управління у XXI ст. // Економіка України. –2005. – № 2. – С. 38-46.
8. Титоренко Г. А. Автоматизированные информационные технологии в экономике. Под. ред. Г.А.Титоренко - М. Компьютер ЮНИТИ, 1998, - 336 с.

Бухно Н.В., Лісова Т.В.
Ніжинський державний університет імені Миколи Гоголя

Програмні засоби для вирівнювання результатів тестування у рамках сучасної теорії тестів IRT

В процесі дослідження проблеми порівняння оцінок різних осіб, що отримані у результаті вимірювання за допомогою різних інструментів (тестів), часто використовують тісно пов'язані між собою терміни linking (зв'язування) та equating (вирівнювання). Термін linking використовують, коли потрібно встановити зв'язок між результатами двох різних тестів, у яких не обов'язково однаковий зміст і рівень складності. Цю проблему, наприклад, доведеться вирішувати приймальним комісіям вищих навчальних закладів під час порівняння сертифікатів ЗНО з різних предметів або різних рівнів. Термін equating використовують для порівняння різних паралельних форм одного тесту, які під час створення вважаються аналогічними за змістом та складністю. Тоді через процес вирівнювання регулюються відмінності у складності і робляться оцінки за різними формами взаємозамінними. Цю процедуру, наприклад, здійснюють під час виставлення балів учасникам ЗНО різних сесій.

Отже, загальноприйнято процес вирівнювання або зв'язування шкал називати linking, а якщо метою вирівнювання є порівняння оцінок учасників, то використовувати термін equating [1].

Якщо розміщуються на одній шкалі результати за паралельними тестами двох груп учасників, розподіли здатностей яких можна вважати однаковими, то говорять про горизонтальне вирівнювання. Інший тип вирівнювання здійснюється у випадку, коли використовуються тести, за якими вимірюють одну характеристику, але тести різних рівнів складності. Наприклад, під час дослідження прогресу у навчанні з певного предмета пропонується кілька різнорівневих тестів для різних років навчання. Тоді процедуру розміщення результатів в одній шкалі з метою їх порівнюваності називають вертикальним вирівнюванням. Статистичні методи горизонтального та вертикального вирівнювання одні і ті самі, але способи їх застосування та інтерпретації дещо відрізняються.

З моменту появи першої процедури вирівнювання для американського армійського тесту, що використовувався після першої світової війни, розроблено багато різних прийомів та методів вирівнювання залежно від мети вирівнювання, процедури тестування та способу збирання відомостей.

Досить ґрунтовно проблему вирівнювання у рамках класичної теорії тестування (КТТ) розглянув Levine R.S. (1955). У КТТ розроблено три основні стратегії вирівнювання [2]: вирівнювання середнього, лінійне та еквіпроцентильне вирівнювання, за яким можна побудувати деяку функцію $Y^* = f(X)$, за якою перетворюються оцінки на шкалі тесту X в еквівалентні оцінки на шкалі тесту Y. За використання усіх методів вирівнювання в КТТ вимагаються серйозні припущення щодо ідентичності розподілів первинних балів та еквівалентності груп учасників, які виконують різні варіанти тесту. Вирівнювання в КТТ лише дозволяє встановити відповідність між балами за різними варіантами тесту і не передбачає побудови спільної шкали.

Lord F.M. (1980) вперше застосував для вирівнювання методи сучасної теорії тестування Item Response Theory (IRT). Застосування методів вирівнювання у рамках IRT дозволяє порівнювати не бали чи процентильні ранги, а об'єктивні оцінки параметрів завдань та учасників, розмістивши їх на одній спільній шкалі. З того часу розроблено багато методів вирівнювання з використанням переваг IRT, які зараз можна умовно поділити на дві групи: методи моментів та методи характеристичних функцій [3].

У роботі [4], яку зараз можна вважати енциклопедією з вирівнювання, зазначається, що будь-яке вирівнювання повинно мати властивості симетричності, прозорості та інваріантності. За застосування методів та підходів до вирівнювання у рамках IRT повністю забезпечується їх виконання.

Процес вирівнювання, незалежно від теоретичного підходу, завжди перебігає через дві фази. Перша фаза – збирання даних для вирівнювання – залежить від способу організації тестування або ж дизайну. Дизайни умовно поділяються на два типи: дизайни єдиної групи, коли можна вважати групи еквівалентними, взятими з однієї вибірки, та дизайни нееквівалентних груп.

Для вирівнювання за методами IRT використовуються три основні дизайни, через які

забезпечують необхідні якірні відомості. Це дизайн рівноваги для єдиної групи, коли вся вибірка випадковим чином ділиться на дві підгрупи, кожна з яких виконує обидва тести, але у різному порядку. Такий дизайн використовується рідко через технічну складність. Дизайн еквівалентних груп, коли кожна з двох груп, відібраних випадковим чином з однієї популяції, виконує свій відмінний тест, є більш привабливим з технічної точки зору. Якірними відомостями тут є лише належність обох груп до однієї популяції. Для більшості методів IRT цього недостатньо, тому для забезпечення додаткових відомостей для процесу вирівнювання вводять до обох тестів спільні завдання, не накладаючи на них особливих умов.

Використання дизайну якірного тесту для нееквівалентних груп передбачає наявність у обох тестах не просто набору спільних завдань, а якірного тесту, який повинен бути «мініверсією» основного тесту і задовільняти ряд вимог: бути придатним для вимірювання того самого конструкту; мати таку ж специфікацію, як основний тест; діапазон складностей повинен максимально перекриватись з усім тестом. Крім того, додатковими вимогами під час використання методів IRT є відповідність завдань якірного тесту обраній моделі та відсутність упередженого функціонування цих завдань у різних групах (DIF). Як правило, кількість якірних завдань складає не менше 20% довжини всього тесту [1].

Наступна фаза процесу вирівнювання – трансформація даних – може відбуватися у різні способи залежно від дизайну, наявності програмного забезпечення та бажаної точності вирівнювання. Перший спосіб полягає у одночасному оцінюванні всіх учасників за обома тестами, його називають конкурентним оцінюванням (CC – concurrent calibration). Даний спосіб може використовуватись із будь-яким дизайном. Необхідно лише правильно підготувати файл з даними, залежно від дизайну, і оцінки параметрів учасників та завдань для обох тестів відразу отримуються у єдиній шкалі. Недоліком такого способу є необхідність опрацьовувати дуже розріджені матриці відповідей великого розміру, що веде до накопичення похибок [4].

За другим підходом (FCIP – fixed common item parameter) тести аналізуються окремо, а для подання результатів проходження одного тесту у шкалі іншого використовуються якірні завдання. Знову ж таки, використання якірних завдань може відбуватися у різний спосіб. Один з них, коли другий тест аналізують із зафіксованими параметрами якірних завдань, отриманими у результаті аналізу першого тесту. Тоді параметри для другої групи автоматично розміщені у шкалі першої групи. Недоліком цього способу є те, що не отримується аналітичний зв'язок між двома шкалами.

Інший спосіб реалізації FCIP полягає у побудові лінійного перетворення

$$\theta_i^X = A\theta_i^Y + B \tag{1}$$

оцінок учасників за тест Y у шкалу тесту X. Тут A та B – деякі дійсні числа, θ_i^X та θ_i^Y – оцінки рівня підготовленості θ i-го учасника у шкалах тесту X та Y відповідно. Існування такого перетворення забезпечується властивостями моделей IRT. Наприклад, для дихотомічних завдань найбільш широко використовується 3-параметрична модель Бірнбаума (3PL), у якій ймовірність того, що i-ий учасник з рівнем підготовки θ_i правильно відповість на j-те завдання складності δ_j з роздільними характеристиками d_j та параметром псевдоугадування c_j визначається за формулою:

$$P_{ij} = P_{ij}\left(\theta_i, \delta_j, d_j, c_j\right) = c_j + \frac{(1-c_j)}{1+\exp(-D\cdot d_j(\theta_i-\delta_j))}. \tag{2}$$

Тут $D=1.7$. Якщо $c_j = 0$, отримаємо 2-парметричну модель (2PL), а якщо ще й $Dd_j = 1$, матимемо 1-параметричну модель або ж модель Раша.

Шкала, у якій можна отримати оцінки латентних параметрів завдань та учасників, є інтервальною з довільним початком та одиницею вимірювання. Як правило, їх вибирають так, щоб середнє значення оцінок латентної характеристики θ_i дорівнювало нулеві, а стандартне відхилення – одиниці для деякої групи опитаних. Зміна одиниці вимірювання та нульової точки інтервальної шкали рівносильна деякому лінійному перетворенню (1). При такому лінійному перетворенні шкали тесту Y в шкалу тесту X параметри завдань d_j та δ_j теж зміняться і матимуть вигляд [3]:

$$d_j^X = \frac{d_j^Y}{A}, \qquad \delta_j^X = A\delta_j^Y + B, \qquad c_j^X = c_j^Y. \tag{3}$$

Сталі A та B можуть визначатись за різними методами. Так, наприклад, за методами моментів

mean/ sigma (Marco, 1977) та *mean/mean* (Loyd & Hoover, 1980) використовують різні статистики якірних завдань.

За методом *mean/sigma* (MS) використовують лише складності якірних завдань, але не враховують їх роздільних характеристик:

$$A = \frac{\sigma(\delta^X)}{\sigma(\delta^Y)},\ B = \mu(\delta^X) - A \cdot \mu(\delta^Y),$$

де $\sigma(\delta)$ та $\mu(\delta)$ – середньоквадратичне відхилення та середнє арифметичне множини складностей якірних завдань у відповідній групі. За методом *mean/mean* (MM) маємо:

$$A = \frac{\mu(d^Y)}{\mu(d^X)},\ B = \mu(\delta^X) - A \cdot \mu(\delta^Y),$$

де $\mu(d)$ – середнє арифметичне параметрів дискримінації якірних завдань. За цими методами отримують різні результати. Інколи перевагу надають MS, аргументуючи тим, що оцінки параметрів складності δ більш стійкі, ніж параметрів d. З іншого боку, за MM використовуються лише середні, які зазвичай більш стійкі, ніж квадратичні відхилення. У зв'язку з цим розроблені різні модифікації методів моментів, що позбавлені вказаних недоліків, наприклад метод *Robust mean/sigma* (Linn, Levine, Hastings, & Wardrop, 1981). Покращення стійкості *A* та *B* у RMS досягається за рахунок введення ваг для кожної пари оцінок якірних завдань:

$$w_i = \frac{1}{\max\left\{\sigma^2(\delta^Y), \sigma^2(\delta^X)\right\}} \text{ та } w_i' = \frac{w_i}{\sum_{j=1}^{k} w_j},$$

де *k* – кількість якірних завдань. Далі *A* та *B* знаходять як за методом MS, використовуючи середнє та квадратичне відхилення уже зважених оцінок $\delta^{Y'} = w_i' \cdot \delta^Y$ та $\delta^{X'} = w_i' \cdot \delta^X$.

На противагу методам моментів було розроблено ряд методів характеристичних функцій (ТСС), де використовують усі параметри завдань, а тому отримують більшу точність. Ідея методів у тому, щоб знайти такі *A* та *B*, за яких різниця між характеристичними функціями якірних завдань чи тестів є мінімальною. За одним методом ТСС (Haebara, 1980) знаходять мінімум функції

$$H_{cr} = \sum_i H(\theta_i),$$

де $H(\theta_i) = \sum_j \left(P_{ij}\left(\theta_i, \delta_j^X, d_j^X\ c_j^X\right) - P_{ij}\left(\theta_i, \delta_j^Y, d_j^Y, c_j^Y\right)\right)^2$ – функція втрат.

Тут підсумовування з індексом *j* проводиться за всіма якірними завданнями, θ_i – точки поділу на осі латентної характеристики, а не рівні підготовки конкретних учасників. За такого підходу забезпечується незалежність перетворення від вибірки опитаних. Можна використовувати також реальні рівні підготовки, або використати щільність нормального розподілу ймовірностей в якості вагових коефіцієнтів до функції втрат.

За іншим методом ТСС (Stocking & Lord, 1983) мінімізують критерій

$$SL_{cr} = \sum_i SL(\theta_i),$$

де $SL(\theta_i) = \left(\sum_j P_{ij}\left(\theta_i, \delta_j^X, d_j^X\ c_j^X\right) - \sum_j P_{ij}\left(\theta_i, \delta_j^Y, d_j^Y, c_j^Y\right)\right)^2$,

а підсумовування з індексом *j* проводиться за всіма завданнями тесту. Тут у дужках маємо різницю характеристичних функцій обох тестів. Цей метод ще називають вирівнюванням істинної оцінки (*true score equating*). Значення *A* та *B*, за яких відповідні критерії досягають мінімуму, знаходять прирівнюванням до нуля частинних похідних за *A* та *B*. За обома методами, незважаючи на різні обчислювальні процедури, отримуються дуже близькі результати. Також ці критерії можна використовувати для порівняння якості вирівнювання за іншими методами, щоб вибрати найкращий.

За використання більшості відомих програмних продуктів (Winsteps, Xcalibre, PARSCALE, jMetrik та ін.), що призначені для моделювання тестів за допомогою математичних моделей, є також певні можливості для здійснення вирівнювання результатів за різними тестами. Практично в усіх

програмах можна реалізувати конкурентне оцінювання (CC), а підхід FCIP лише у вигляді оцінювання іншої групи із зафіксованими параметрами якірних завдань.

Як правило, в таких програмах передбачено широкий набір послуг не тільки щодо моделювання, а й для дослідження різних супутніх проблем, тому вони є переважно комерційними і їх використання у навчальних та дослідницьких цілях не завжди можливе. У вільному доступі можна знайти, наприклад, програму jMetrik, яку її автор, професор університету штату Вірджинія J.P. Meyer, розповсюджує безкоштовно з 2009 року, але за супровідну документацію та on-line консультації потрібно платити. jMetrik є Java-додатком, що функціонує з операційними системами Windows, Linux, постійно вдосконалюється, але на сьогодні в ньому обмежений набір моделей та можливостей використання.

Група ентузіастів-дослідників із психометрії «Measured Progress» у співпраці з професором статистики університету штату Іллінойс W.F. Stout мають намір у майбутньому створити набір вільно доступних програм для проведення психометричних досліджень з відкритим кодом [5]. А поки що студенти можуть знайомитися з можливостями використання комерційних програм лише за їх безкоштовними версіями з обмеженнями, як правило, на кількість учасників та завдань.

Більше у вільному доступі існує програм, які призначені безпосередньо для процедури вирівнювання (GGUMLINK, IPLINK, EQUATE, IRTEQ та ін.). У них реалізовано ширший набір методів FCIP з побудовою лінійного перетворення, але вони застосовні до уже отриманих у будь-який спосіб оцінок параметрів завдань та учасників за обома тестами. Такі програми не відрізняються універсальністю і не завжди оснащені зручним інтерфейсом. Так, у програмі GGUMLINK (Roberts & Huang) реалізовані усі п’ять вище згаданих методів вирівнювання. Вона описана мовою FORTRAN, орієнтована на Windows, під час роботи з нею потрібна участь користувача для подання текстових команд [6].

Програма IPLINK (Lee & Oshima) написана в Turbo C++ і функціонує на базі Windows. Тут реалізовані методи TCC, допускається введення більше однієї пари файлів вхідних даних. У [1] для демонстрації процесу вирівнювання за методами TCC використовується програма EQUATE (Baker), але вона створена для DOS. Більш потужним та зручним засобом для вирівнювання є Windows-додаток IRTEQ (Han), остання версія якого випущена у 2011 р. та розміщена на сайті Центру освітніх вимірювань університету штату Массачусетс [7]. Тут розглянемо можливості використання даної програми та комерційної Winsteps [8].

У програмі Winsteps (Linacre & Wright) реалізовані моделі сімейства Раша для дихотомічних та політомічних завдань, тому за належної відповідності даних побудованій моделі для вирівнювання двох тестів достатньо знати лише сталу *B*, оскільки у всіх кривих однаковий нахил і $A=1$. Але тут не пропонується пряме обчислення *B*, натомість є кілька можливих варіантів для процедури вирівнювання.

Найпростіше реалізувати конкурентне оцінювання в одному файлі відповідно до дизайну із спільними завданнями чи учасниками (рис. 1). Але ця процедура не позбавлена недоліків.

По-перше, відповіді на ті завдання, що не пропонувались одній із груп, повинні виключатися з аналізу. Для цього у контрольному файлі можна вказати змінну MISSING= –1. Але тоді не адмініструватимуться усі пропуски. Якщо вважати неправильними відповідями ті пропуски, що допущені учасниками під час проходження свого тесту, то потрібно ці пропуски замінити нулями.

По-друге, якірні завдання можуть суттєво по-різному функціонувати у групах, що спотворює процес вирівнювання. Щоб не потрапити в таку ситуацію, необхідно у контрольному файлі вказати змінні CUTLO= або CUTHI=, щоб з аналізу вилучались ті завдання, параметри яких відхиляються більше за вказані межі. У будь-якому випадку варто почати вирівнювання з аналізу параметрів якірних завдань та їх розсіяння навколо прямої з нахилом $k = \sigma(\delta^Y)/\sigma(\delta^X)$.

За другим способом спочатку аналізують результати першої групи та у окремому файлі зберігають параметри якірних завдань чи учасників. Перед початком аналізу даних другої групи на запит «Extra Specification» за допомогою змінних IAFILE= або IPFILE= вказуємо ім’я збереженого файлу. У вихідних таблицях матимемо ще й повідомлення про зміщення параметрів якірних завдань.

За третім способом з окремих аналізів обчислюємо

$$A = \sigma(\delta^X)/\sigma(\delta^Y) \text{ та } B = \mu(\delta^X) - \mu(\delta^Y).$$

Далі знову аналізуємо другий тест, включивши у контрольний файл змінні USCALE=*A* та UMEAN=*B*, чим задається новий початок та одиниця системи відліку.

Програма IRTEQ оснащена зручним інтерфейсом, її використання не потребує введення додаткових команд. Розглянемо детальніше можливості її використання для вирівнювання.

На рис. 2 зображено одне з двох головних вікон програми.

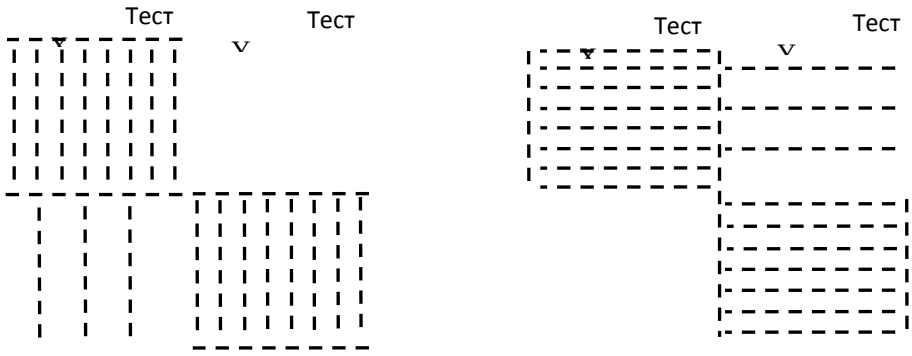

Рис. 1.

Цифрами відмічено кроки, які необхідно здійснити для визначення параметрів аналізу:

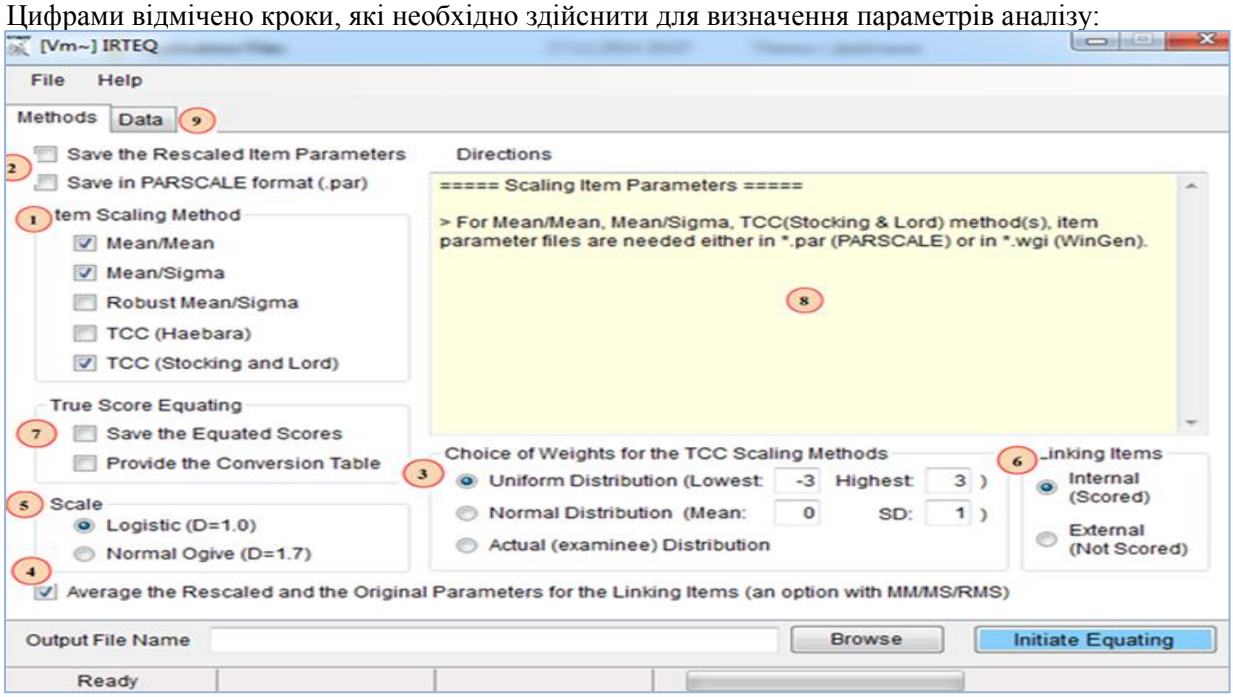

Рис. 2.

1. Вибрати принаймні один метод для зміни масштабу тестових завдань.

2. (Необов'язково) Встановити відповідний прапорець, якщо потрібно зберегти параметри завдань після зміни масштабу у форматі WinGen (*.wgi) або у форматі PARSCALE (*.par).

3. Якщо у пункті 1 обрано методи характеристичних функцій (TCC), то необхідно обрати один з розподілів θ, що буде використовуватись під час мінімізації функції втрат. Якщо нема чіткої впевненості у вигляді такого розподілу, варто обирати дискретний «Actual (examinee) distribution».

4. Якщо у пункті 1 обрано методи MM, MS чи RMS, то можна отримати усереднені параметри якірних завдань, вибравши цю опцію.

5. Вибрати шкалу логітів чи пробітів. Для моделей Раша, як правило, обирають D = 1.0, а для 2PL чи 3PL обирають D = 1.7.

6. Вибрати варіант використання якірного тесту. За варіантом «Internal» відповіді учасників на питання якірного тесту враховуються у його загальну оцінку, а за варіантом «External» завдання якірного тесту не враховуються у загальну оцінку, а використовуються лише для вирівнювання.

7. (Необов'язково) Встановити прапорець «Save Equated Scores», якщо потрібно зберегти результати тесту 2 у шкалі тесту 1 у форматі WinGen (*.wge) або PARSCALE (*.sco). Вибираючи «Provide Conversion Table», отримаємо таблицю перетворених “сирих” балів. У поточній версії програми за використання моделей з параметром угадування перетворена тестова оцінка не може бути нижчою за поріг угадування (суми всіх c_j).

Наступним етапом є введення даних і запуск програми (рис.3).

1. Завантажити параметри завдань тесту 1, натиснувши «Open File». Дані можуть бути

заздалегідь отримані у будь-якій програмі, але підготовлені у форматі WinGen (.wge) або PARSCALE (.par).

2. Завантажити параметри завдань тесту 2, які потрібно перетворити у шкалу тесту 1. Переглянути відкриті параметри завдань у списку нижче.

3. Список якірних завдань може бути сформований вручну або введений із раніше створеного файлу у форматі *.lkl. Для створення списку вручну вказати якірні завдання в обох тестах та натиснути на кнопку «Add». Для вилучення пари якірних завдань використовують «Remove».

4. Натиснувши «abc-Plot», можна у окремому вікні (рис.4) переглянути взаємне розташування параметрів якірних завдань. Якщо завдання у обох групах функціонують однаково, їх параметри не повинні бути дуже розсіяні навколо бісектриси. У іншому вікні «Plot TCC» доступні для перегляду графіки характеристичних кривих обох тестів до вирівнювання. TCC для масштабованого тесту 2 буде доступна після виконання програми.

5. Якщо на першому етапі було обрано опції «Save Equated Scores» або «Actual Distribution» для TCC методу, необхідно завантажити файл з оцінками учасників за тест 2.

6, 7. Вказати ім'я вихідного файлу та натиснути «Initiate Equating».

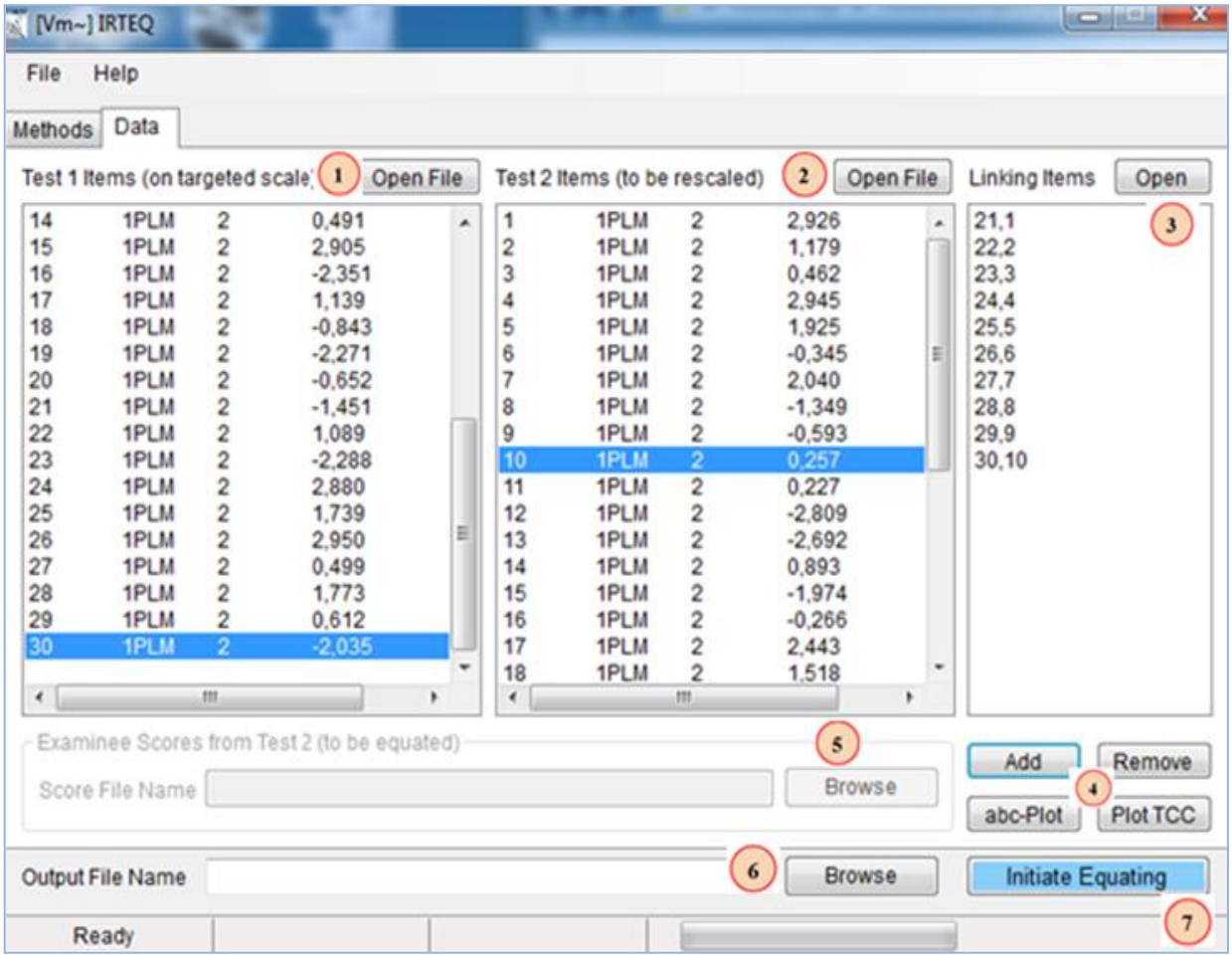

Рис. 3.

Після завершення процесу вихідний файл з розширенням *.OUT, фрагмент якого зображено на рис. 5, відкривається автоматично. В ньому містяться характеристики параметрів якірних завдань, що використовувались у процесі вирівнювання, та обчислені за різними методами коефіцієнти *A* та *B*.

Програма IRTEQ придатна для підтримки більшості відомих одновимірних моделей тестів з дихотомічними та політомічними завданнями, а тому її можна використовувати для вирівнювання результатів двох тестів будь-яких реальних досліджень. Для цього потрібно лише підготувати дані з параметрами усіх завдань, отримані довільним доступним способом, у форматах даної програми.

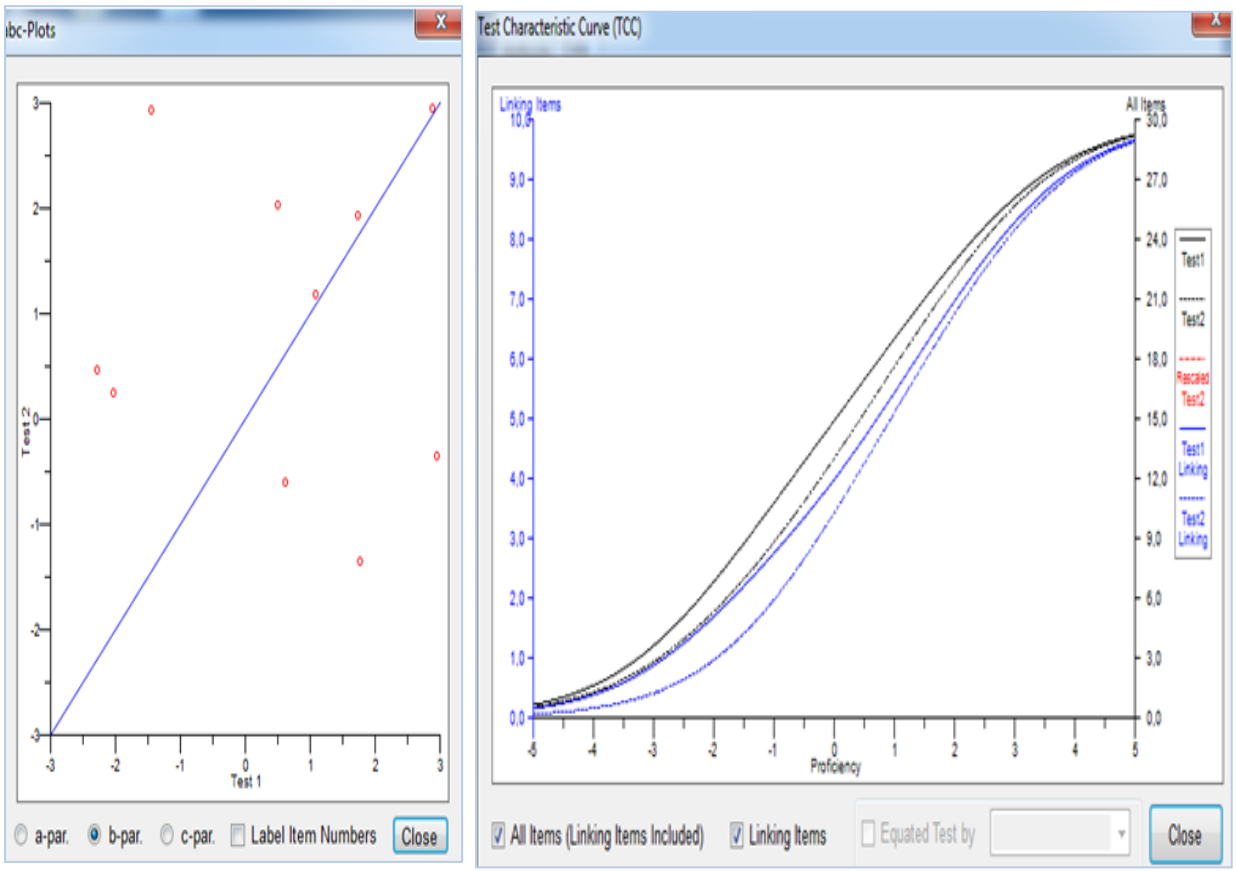

Рис. 4.

```
Equating coefficient A by Mean-Mean method:                1,00
Equating coefficient B by Mean-Mean method:                -0,37

Equating coefficient A by Mean-Sigma method:               1,28
Equating coefficient B by Mean-Sigma method:               -0,63

Equating coefficient A by TCC (Haebara) method:                    0,50
Equating coefficient B by TCC (Haebara) method:                    0,13
Minimized Loss Function Value by TCC (Haebara) method:             0,91738

Equating coefficient A by TCC (Stocking & Lord) method:            1,30
Equating coefficient B by TCC (Stocking & Lord) method:            -0,60
Minimized Loss Function Value by TCC (Stocking & Lord) method:     0,01075
```

Рис. 5.

Крім того, вона дуже зручна під час вивчення методів вирівнювання, оскільки формати всіх файлів сумісні з відповідними форматами програми WinGen того самого автора, яка призначена для генерування даних у широкому діапазоні моделей тестів.

Список використаних джерел

1. De Ayala R.J. The theory and practice of Item Response Theory / Rafael J. de Ayala. – N.Y., L.: The Guilford Press, 2009. – 448 p.

2. Крокер Л. Введение в классическую и современную теорию тестов / Линда Крокер, Джеймс Алгина. – М.: Логос, 2010. – 668 с.

3. Лісова Т.В. До проблеми вирівнювання результатів тестування у рамках сучасної теорії IRT / Т.В. Лісова // Гуманітарний вісник – Додаток 4 до Вип. 31, Том ІІІ (11): Тематичний випуск «Міжнародні Челпанівські психолого-педагогічні читання». – К.: Гнозис, 2014. – С. 368-375.

4. Kolen M.J. Test Equating, Scaling, and Linking: Methods and Practices / Michael J. Kolen, Robert L. Brennan (2nd ed.). – N.Y.: Springer, 2004. – 549 p.

5. http://psychometrictools.measuredprogress.org/home

6. http://www.psychology.gatech.edu/Unfolding/FreeSoftware.html

7. http://www.umass.edu/remp/main_software.html

8. http://www.winsteps.com