**ЛІСОВА Т.В.,**
кандидат фізико-математичних
наук, доцент,
кафедра прикладної математики,
інформатики та освітніх
вимірювань,
Ніжинський державний
університет імені Миколи Гоголя,
м. Ніжин

# ДО ПРОБЛЕМИ ВИРІВНЮВАННЯ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ У РАМКАХ СУЧАСНОЇ ТЕОРІЇ IRT

*У статті розглядаються основні методи вирівнювання результатів тестування у рамках одновимірних моделей сучасної теорії IRT, їх властивості та умови застосування.*

***Ключові слова:*** *зв'язування, вирівнювання, дизайн вирівнювання, інтервальна шкала, моделі IRT, якірні завдання.*

*В статье рассматриваются основные методы выравнивания результатов тестирования в рамках одномерных моделей современной теории IRT, их свойства и условия применения.*

***Ключевые слова:*** *связывание, выравнивание, дизайн выравнивания, интервальная шкала, модели IRT, якорные задания.*

*This paper describes the main methods of aligning the test results within a one-dimensional model of IRT, their properties and conditions of use.*

***Key words:*** *linking, equating, equating design, interval scale, model of IRT, anchor items.*

**Постановка проблеми та її актуальність.** Проблема вирівнювання результатів тестування виникає тоді, коли різні учасники оцінюються за допомогою різних інструментів, що призначені для вимірювання однієї і тієї ж характеристики. Наприклад, при використанні різних варіантів одного тесту у різних групах учасників виникає питання про рівноцінність балів учасників, отриманих за цими варіантами. Хоча розробники тестів намагаються створювати максимально паралельні варіанти, на практиці отримати повністю однакові їх характеристики не вдається. Для того, щоб бали за різними варіантами можна було порівнювати, необхідно провести процедуру вирівнювання, яка дозволяє встановити зв'язок між балами учасників за різними варіантами тесту та розмістити їх на одній спільній шкалі. У даному випадку цю процедуру називають *горизонтальним вирівнюванням*. Саме горизонтальне вирівнювання здійснюється для виставлення остаточних балів учасників ЗНО різних сесій. Аналогічна проблема виникає при порівнюванні складностей завдань, отриманих у різних групах тестованих, з метою їх розміщення на спільній шкалі для створення банку тестових завдань. І першу, і другу проблеми необхідно вирішувати при проведенні адаптивного тестування, коли кожен учасник отримує свій власний варіант тесту, сформований із завдань такої складності, що відповідають виключно його рівню здатності.

Інший тип вирівнювання здійснюється у випадку, коли використовуються тести, що вимірюють одну характеристику, але мають різні рівні складності. Тоді відповідну процедуру називають *вертикальним вирівнюванням*. Наприклад, при дослідженні прогресу у навчанні (моніторингу досягнень) розробляються кілька різнорівневих тестів. Тест нижчого рівня пропонується на початку навчання, а далі пропонуються тести наступних рівнів – після завершення чергового етапу навчання. Виникає необхідність розмістити оцінки на кожному рівні у одній шкалі, щоб мати можливість оцінити ступінь прогресу. Потребує вертикального вирівнювання і ситуація, коли для слабких учасників тести вищого рівня можуть виявитися занадто складними, що не забезпечить якісного вимірювання. Тоді таким учасникам дають змогу пройти тест нижчого рівня, а для інтерпретації результатів та можливості їх порівнюваності оцінку переводять у шкалу тесту вищого рівня. Аналогічно для «вундеркіндів», які легко виконують всі завдання тесту даного рівня і отримують просто максимальну оцінку без належного вимірювання їх рівня здатності, можна пропонувати тести вищого рівня, а оцінку подати у загальній шкалі для всіх учасників.

В іншомовній літературі при дослідженні даних проблем широко використовуються тісно пов'язані між собою терміни *linking* (зв'язування) та *equating* (вирівнювання). При цьому процес вирівнювання або зв'язування шкал називають *linking*, а якщо метою

вирівнювання є порівняння оцінок учасників, використовують термін *equating* [5, ст. 306].

В класичній теорії тестування (КТТ) розроблено три основні стратегії вирівнювання: вирівнювання середнього, лінійне та еквіпроцентильне вирівнювання, які дозволяють побудувати деяку функцію $Y^* = f(X)$, що перетворює оцінки на шкалі тесту *X* в еквівалентні оцінки на шкалі тесту *Y*. При лінійному вирівнюванні ця функція лінійна. Дану стратегію використовують, коли розподіли оцінок за тестами *X* та *Y* відрізняються лише середніми та стандартними відхиленнями. Якщо стандартні відхилення для обох тестів однакові, використовують частинний випадок лінійного вирівнювання – вирівнювання середнього. При еквіпроцентильному вирівнюванні не накладаються такі обмеження на розподіли, а функція $f(X)$ може бути нелінійна. Тут оцінки за два варіанти тесту вважаються еквівалентними, якщо їм відповідають однакові процентильні ранги. Усі методи вирівнювання в КТТ вимагають серйозних припущень щодо ідентичності розподілів первинних балів та еквівалентності груп учасників, що виконують різні варіанти тесту. Вирівнювання в КТТ лише дозволяє встановити відповідність між балами за різними варіантами тесту і не передбачає побудови спільної шкали.

Методи вирівнювання у рамках сучасної теорії тестування Item Response Theory (IRT) дозволяють порівнювати не бали чи процентильні ранги, а об'єктивні оцінки параметрів завдань та учасників, розмістивши їх при цьому на одній спільній шкалі. Вирівнювання у рамках IRT засноване швидше на абстракції, ніж на реальних статистичних показниках, тому має ряд сильних припущень, що можуть не виконуватись при реальному тестуванні. Таке вирівнювання є складнішим у концептуальному та процедурному смислі, однак має ряд переваг. Зокрема, воно є більш гнучким у виборі плану дій для ув'язки різних тестових форм, що дуже корисно при адаптивному тестуванні, може використовуватись у випадках, коли традиційні методи не працюють, наприклад, при вирівнюванні наборів завдань.

**Аналіз наукових праць, присвячених розв'язанню проблеми.** Метод лінійної регресії був одним із перших методів, що використовувався для передбачення результатів одного тесту за відомими результатами іншого. Thorndike E.L. та Otis A.S. (1922) вперше вказали на різницю між передбаченням та вирівнюванням, як ми це зараз розуміємо. Далі у роботах Flanagan J.C. (1939, 1951), Lord F.M. (1950, 1955, 1982), Angoff W.H. (1971), Mislevy R.J. (1992) та Linn R.L. (1993) досліджувались різні аспекти цієї відмінності, вводились поняття еквівалентних оцінок, паралельних тестів тощо. Thurstone L.L. (1925, 1938) запропонував метод побудови вертикальної шкали для вирівнювання тестів різної складності. Термін *equating* стосовно порівнюваності балів вперше використав Kelley T.L. (1923). Першим відомим тестом, для якого розроблялись процедури вирівнювання, був американський армійський тест, що використовувався після першої світової війни у двох варіантах «Alpha» – для тих, хто володіє англійською, та «Beta» – для тих, хто не володіє. З тих пір розроблено багато різних прийомів та методів вирівнювання залежно від процедури тестування та способу збору інформації.

Проблему вирівнювання у рамках класичної теорії тестування виклав Levine R.S. (1955), а Lord F.M. (1980) вперше застосував для вирівнювання методи IRT. Теоретичному обґрунтуванню проблеми вирівнювання тестів присвячені роботи Morris C.N. (1982), Hanson B.A. (1991), van der Linden W.J. (2000). Практичні методи для побудови лінійного перетворення з метою вирівнювання результатів у рамках IRT запропоновані Marco G.L. (1977), Loyd B.H. and Hoover H.D. (1980), Haebara T. (1980), Stocking M. and Lord F.M. (1983).

Багато робіт цих та інших авторів присвячено проблемі точності вирівнювання.

Практичним посібником для проведення процедури вирівнювання усіма можливими методами можна вважати [6], у роботі [5] розглядаються основні методи у рамках IRT, а у [7] надається енциклопедична довідка щодо усіх можливих проблем, пов'язаних з вирівнюванням. У роботі [1] представлено застосування багатопараметричної моделі Раша для вирівнювання результатів тестів, які пов'язані не спільними завданнями, а спільними експертами, що оцінюють різні варіанти.

**Постановка завдання, цілі статті.** Будь-яке вирівнювання, незалежно від підходу, повинно мати певні загальновизнані властивості: властивість *симетричності* означає, що функція переведення результатів зі шкали *X* у шкалу *Y* повинна бути оберненою до функції, що переводить результати з *Y* в *X*; властивість *справедливості* означає, що розподіли балів до та після вирівнювання повинні зберігати форму; властивість *інваріантності* означає, що функція вирівнювання не повинна залежати від вибірки, що бере участь у вирівнюванні. У роботі планується розглянути основні методи вирівнювання результатів тестування у рамках сучасної теорії IRT, які мають всі вказані властивості, визначити умови їх застосування.

**Виклад основного матеріалу.** Серед одновимірних моделей IRT найбільш широко використовується 3-параметрична модель Бірнбаума, у якій ймовірність того, що *i*–ий учасник з рівнем підготовки $\theta_i$ вірно відповість на дихотомічне *j* –те завдання з параметрами $c_j$ (складність), $c_j$ (роздільна здатність) та $c_j$ (псевдоугадування) визначається формулою:

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

Тут $D = 1.7$. Якщо $c_j = 0$, отримаємо 2-парметричну модель, а якщо ще й $Dd_j = 1$, матимемо 1-параметричну модель або ж модель Раша. Відомості про основні властивості параметрів цих моделей, характеристичні та інформаційні функції завдань та тесту, методи побудови оцінок даних параметрів можна знайти у роботах [2, 3, 5].

Шкала, у якій вимірюються латентні характеристики завдань та учасників, є інтервальною і має довільний початок та одиницю вимірювання. Як правило, їх вибирають так, щоб середнє значення оцінок латентної характеристики $\theta_i$ дорівнювало нулеві, а стандартне відхилення – одиниці для деякої групи опитаних. Уявимо, що проводиться опитування у двох неекві валентних групах за допомогою різних варіантів тесту: група 1 виконує варіант *X*, а група 2 – варіант *Y*. У результаті оцінювання отримаємо дві шкали зі своїми нулями та одиницями. Ці шкали могли б збігатися у випадку, коли групи повністю еквівалентні, а варіанти строго паралельні. На практиці таке буває дуже рідко, тому необхідне перетворення шкал таким чином, щоб на спільній шкалі були відображені відмінності між групами, а параметри завдань та учасників можна було б порівнювати.

Зміна одиниці вимірювання та нульової точки інтервальної шкали рівносильна деякому лінійному перетворенню

$$\theta_i^X = A\theta_i^Y + B, \quad (2)$$

де *A* та *B* – дійсні числа, $\theta_i^Y$ та $\theta_i^Y$ – оцінки параметра $d_j$ *i*-го учасника у шкалі тесту *X* та *Y* відповідно. При перетворенні шкали тесту *Y* в шкалу тесту *X* параметри завдань $d_j$ та $\delta_j$ теж зміняться, оскільки в обох шкалах одна й та ж ймовірність правильної відповіді має вигляд (1), але з різними параметрами:

$$c_j^X + \frac{(1-c_j^X)}{1+\exp(-D\cdot d_j^X(\theta_i^X-\delta_j^X))} =$$

$$= c_j^Y + \frac{(1-c_j^Y)}{1+\exp(-D\cdot d_j^Y(\theta_i^Y-\delta_j^Y))}.$$

Підставляючи сюди вираз (2), матимемо у новій системі координат

$$d_j^X = \frac{d_j^Y}{A},\ \delta_j^X = A\delta_j^Y + B,\ c_j^X = c_j^Y. \quad (3)$$

Сталі *A* та *B* лінійного перетворення (2) можна визначити за відомою додатковою інформацією про значення параметрів мінімум двох учасників або двох завдань у обох шкалах. Наприклад, якщо два учасники $i$ та $i^*$ виконають обидва варіанти $X$ та $Y$ (або учасники обох груп виконають два завдання $j$ та $j^*$), ми отримаємо параметри $\theta_i^X$, $\theta_i^Y$, $\theta_i^Y$ та $\theta_{i^*}^Y$ (або $\delta_{j^*}^X$, $\delta_{j^*}^X$, $\delta_j^Y$ та $\delta_{j^*}^Y$), пов'язані між собою співвідношенням (2), то легко знайти

$$A = \frac{\theta_i^X - \theta_{i^*}^X}{\theta_i^Y - \theta_{i^*}^Y} = \frac{\delta_j^X - \delta_{j^*}^X}{\delta_j^Y - \delta_{j^*}^Y} = \frac{d_j^Y}{d_j^X},$$

$$B = \theta_i^X - A\cdot\theta_i^Y = \delta_j^X - A\cdot\delta_j^Y.$$

Зауважимо, що аналогічно можна побудувати лінійне перетворення оцінок шкали $X$ у шкалу $Y$, тобто таке перетворення є симетричним. Крім того, моделі IRT дозволяють отримати оцінки параметрів учасників та завдань, які не залежать від контингенту опитаних, а тому перетворення теж вільне від такої залежності і є унікальним. Усі ці особливості моделей IRT дозволяють успішно їх використовувати для побудови єдиних шкал.

Процес вирівнювання, незалежно від теоретичного підходу, завжди має дві фази. Перша фаза – збір 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 методами IRT використовуються три основні дизайни. Це дизайн рівноваги для єдиної групи (*counterbalanced random-groups design*), коли вся вибірка випадковим чином ділиться на дві підгрупи, кожна з яких виконує обидва тести, але у різному порядку. Такий дизайн дає багато спільної (якірної) інформації, що полегшує процес вирівнювання, однак використовується рідко через значні затрати часу на адміністрування досить довгих тестів.

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. Якірною інформацією тут є лише належність обох груп до однієї популяції. Для більшості методів IRT цього не достатньо, тому для забезпечення додаткової інформації для процесу вирівнювання вводять до обох тестів спільні завдання (*common-item random groups design*), не накладаючи на них особливих умов.

Дизайн якірного тесту для нееквівалентних груп (*common-item nonequivalent groups design*) передбачає наявність у обох тестах набору спільних завдань – якірного тесту, який повинен бути «мініверсією» основного тесту і задовольняти ряд вимог. Він повинен вимірювати то самий конструкт і мати таку ж специфікацію, як основний тест. Діапазон складностей завдань якірного тесту повинен максимально перекриватись з усім тестом. Крім того, додатковими вимогами при використанні методів IRT є належна відповідність завдань якірного тесту обраній моделі (model fit) та відсутність упередженого функціонування цих завдань у різних групах (DIF). Як правило, довжина якірного тесту складає принаймні 20 завдань або не менше 20% довжини

всього тесту [4 - 6]. Можливі два варіанти використання дизайну якірного тесту. За першим варіантом (*internal common items*) відповіді учасників на питання якірного тесту враховуються у його загальну оцінку, а самі питання можуть бути розподілені по всій довжині тесту. За іншим варіантом (*external common items*) завдання якірного тесту пропонуються після або перед основним тестом і відповіді на них не враховуються у загальну оцінку.

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. Може використовуватись із будь-яким дизайном. Необхідно лише правильно підготувати файл з даними залежно від дизайну і оцінки параметрів учасників та завдань зразу отримуються у єдиній шкалі. При цьому ті завдання, що не виконувалися, не адмініструються у відповідних групах, тому необхідно розрізняти пропущені завдання всередині виконуваного тесту та такі, що не пропонувалися. У різних програмах для цього передбачені різні засоби. Недоліком такого способу є необхідність обробляти дуже розріджені матриці відповідей великого розміру, що веде до накопичення похибок.

За другим підходом аналізуються результати у одній із груп та визначаються параметри якірних завдань (*anchor items*). Аналіз для другої групи здійснюють, зафіксувавши вже знайдені параметри якірних завдань. У більшості програмних продуктів така можливість передбачена. Таким чином, параметри для другої групи автоматично будуть розміщені у шкалі першої групи. Недоліком цих підходів є те, що вони не дають аналітичного зв'язку між двома шкалами.

Третій підхід полягає у побудові аналітичного лінійного перетворення (2) за допомогою результатів, отриманих у кожній групі окремо. Сталі *A* та *B* можуть знаходитись різними методами. Так, наприклад, методи моментів *mean/ sigma* (Marco, 1977) та *mean/ mean* (Loyd and Hoover, 1980) використовують різні статистики якірних завдань.

Метод *mean/sigma* використовує лише складності якірних завдань, але не враховує їх роздільну здатність:

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

де $\sigma(\delta)$ та $\mu(\delta)$ – середньоквадратичне відхилення та середнє арифметичне множини складностей якірних завдань у відповідній групі. За методом *mean/mean* 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. Ці методи дають різні результати. Інколи перевагу надають *mean/ sigma*, аргументуючи тим, що оцінки параметрів складності $d$ більш стійкі, ніж параметрів $d$ . З іншого боку, *mean/mean* використовує лише середні, які зазвичай більш стійкі, ніж квадратичні відхилення. У зв'язку з цим розроблені різні модифікації методів моментів, які не мають вказаних недоліків, а також різні статистики, що дозволяють серед кількох методів обрати кращий.

На противагу методам моментів були розроблені методи *характеристичних функцій* , які використовують усі параметри завдань, а тому забезпечують більшу точність. Ідея методів у тому, щоб знайти такі *A* та *B* при яких різниця між характеристичними функціями якірних завдань чи тестів є мінімальною. За методом Haebara (1980) потрібно знаходити мінімум функції $H_{cr} = \sum_i H(\theta_i)$, де

$$H(\theta_i) = \sum_j \left( P_{ij}\left(\theta_i, \delta_j^X, d_j^X\ c_j^X\right) -\right.$$

$- P_{ij}\left(\theta_i, \delta_j^Y, d_j^Y, c_j^Y\right)\Big)^2$

– функція втрат. Тут сумування з індексом $j$ проводиться по всіх якірних завданнях, $\theta_i$ – точки розбиття на осі латентної характеристики, а не рівні підготовки конкретних учасників. Такий підхід (має назву *uniform*) забезпечує незалежність перетворення від вибірки опитаних. Можна використовувати реальні рівні підготовки (*actual*), або використати щільність нормального розподілу в якості вагових коефіцієнтів до функції втрат (*normal*). За методом Stocking and Lord (1983) мінімізують критерій $SL_{cr} = \sum_i SL(\theta_i)$,

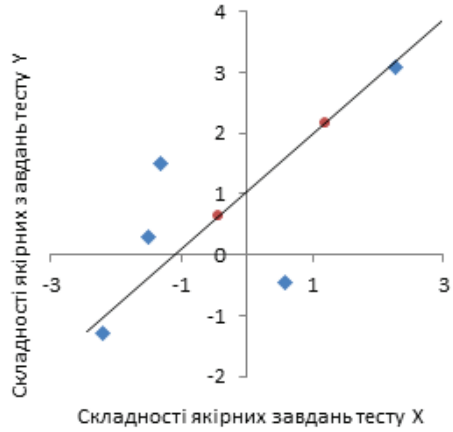
де

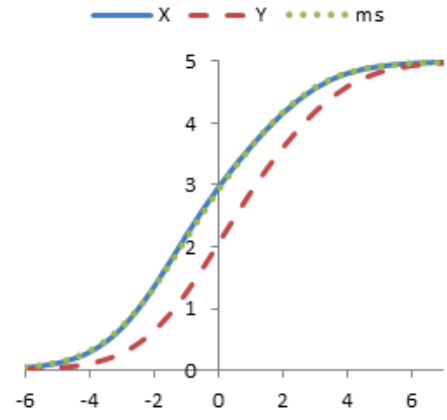
$$SL(\theta_i) = \left( \sum_j P_{ij}\left(\theta_i, \delta_j^X, d_j^X\ c_j^X\right) - \right.$$

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

а сумування з індексом $j$ проводиться по всіх завданнях тесту. Тут у дужках маємо різницю характеристичних функцій обох тестів. Цей метод ще називають вирівнюванням істинної оцінки (*true score equating*). Значення $A$ та $B$, при яких відповідні критерії досягають мінімуму, знаходять прирівнюванням до 0 частинних похідних по $A$ та $B$. Обидва методи, незважаючи на різні обчислювальні процедури, дають дуже близькі результати. Обидва критерії можна використовувати для порівняння якості вирівнювання іншими методами, щоб вибрати найкращий.

Для прикладу вирівняємо результати двох тестів по 15 завдань (5 спільних з $\mu(\delta^X) = -0.44$; $\mu(\delta^Y) = 0.63$; $\sigma(\delta^X) = 1.63$; $\sigma(\delta^Y) = 1.54$), згенерованих у програмі WinGen, методом *mean/sigma*, використовуючи модель Раша: $A = 1.06$, $B = -1.1$. Параметри спільних завдань не сильно розсіяні навколо прямої, що проходить через середні з нахилом $A = 1.06$, а отже вирівнювання має бути ефективним. Дійсно, після перетворення $1.06 * \theta - 1.1$ характеристичні криві якірного тесту зливаються. Отже, оцінки всіх учасників за тест $Y$ потрібно перерахувати за формулою $1.06 * \theta_i^Y - 1.1$.




**Висновки та перспективи подальших розробок дослідження проблеми.** У даній роботі розглянуто методи вирівнювання результатів тестування у рамках деяких одновимірних моделей IRT для тестів з дихотомічними завданнями, проведена їх систематизація та класифікація. Усі методи задовольняють загальні властивості вирівнювання. Крім того, важливою умовою застосування методів IRT є відповідність емпіричних даних обраній моделі. Далі необхідно розглянути методи вирівнювання тестів з політо-

мічними завданнями та комбінованих тестів з використанням одновимірних та багатовимірних моделей IRT, методи вертикального вирівнювання та доступні для цього програмні засоби.

***СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:***


1. *Карданова Е.Ю. Выравнивание показателей в случае експертного оценивания заданий / Елена Карданова // Педагогика и психология. Известия Томского политехнического университета, Т. 310, № 3. – 2007. – С. 233-237.*
2. *Крокер Л. Введение в классическую и современную теорию тестов / Линда Крокер, Джеймс Алгина. – М.: Логос, 2010. – 668 с.*
3. *Нейман Ю.М. Введение в теорию моделирования и параметризации педагогических тестов / Ю.М. Нейман, В.А. Хлебников. – М.: Прометей, 2000. – 168 с.*
4. *Angoff W.H. Scales, norms and equivalent scores / In R.L. Thorndike (Ed.), Educational measurement (2nd ed.). – Washington, DC: American Council on Education, 1971. – P. 508–600.*
5. *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.*
6. *Kolen M.J. Test Equating, Scaling, and Linking: Methods and Practices / Michael J. Kolen, Rjbert L. Brennan (2nd ed.). – N.Y.: Springer, 2004. – 549 p.*
7. *Holland P.W. Linking and equating / Paul W. Holland, Neil J. Dorans, In R.L. Brennan (Ed.), Educational measurement (4nd ed.). – Westport, CT: Praeger, 2006. – P. 187–220.*