УДК 001.891:371.263

М. Є. СІНИЦЬКИЙ,
кандидат фіз.-мат. наук, доцент,
Національна академія статистики, обліку та аудиту

Статистичні інструменти вимірювання якості освіти. Частина 3. Класичний підхід

Представлено огляд статистичних основ тестології. Описано базові задачі, математичні моделі та основні розрахункові формули сучасної (IRT) теорії тестування, що дозволяє зі статистичних позицій оцінити правильність побудови, роздільну здатність, стандартну помилку та надійність тестових вимірювань.

Ключові слова: *класична теорія тестування, випробовувані, оцінювання, тест, IRT-модель, PL-модель, PCM-модель, шкала Раша, логіт-оцінка.*

У попередніх публікаціях [1; 2] було розглянуто елементи класичної теорії тестування (CTT). Головним недоліком CTT вважають залежність результатів вимірювання знань випробовуваних від характеристик тестових завдань. Проблема полягає в тому, що первинний бал істотно залежить від труднощів завдань тесту, причому сама трудність тесту визначається всім контингентом випробовуваних. Крім того, первинні бали є нелінійними відносно рівня підготовленості. Вони є лише індикатором підготовленості випробовуваних, а не її мірою.

Все це суттєво ускладнює порівняння результатів тестів, що відрізняються завданнями (оскільки створювати дійсно паралельні тести майже неможливо, не кажучи про полідисциплінарні тести або такі, що надаються різним групам. Цю та інші проблеми CTT багато в чому вирішує Item Response Theory (IRT) [3; 5], або у непрямому перекладі теорія моделювання та параметризації тестів. Основний предмет застосування IRT-оцінка ймовірності правильної відповіді випробовуваних на завдання різної трудності.

Пик розвитку цієї галузі досліджень (Л. Такер, Л. Гутман, П. Лазарсфельд, М. Новік, Ф. Лорд, Г. Раш та інш.) припадає на 50–70-х роках минулого сторіччя. Певною віхою на шляху становлення IRT стала модель, запропонована датським математиком Г. Рашем у 1960 р. Підхід Г. Раша ґрунтується на припущенні, що ймовірність (P) правильного виконання завдання[1] є функцією (її називають функцією успіху) принаймні двох латентних змінних [6]:

$$P = P(u, t),\ P \in [0,1] \tag{1}$$

де u – рівень підготовленості випробовуваного; $u \in [0, \infty]$;

t – рівень труднощі (трудності) завдання тесту; $t \in [0, \infty]$.

За Г. Рашем для забезпечення об'єктивності порівняння випробовуваних повинні виконуватись рівняння: $\frac{u_1}{u_2} = \frac{t_1}{t_2} = \ldots$, тоді ймовірності правильної відповіді випробовуваним з рівнем підготовленості u_i на завдання труднощі t_i має співпадати з ймовірністю правильної відповіді випробовуваним з рівнем підготовленості u_j на завдання труднощі t_j. Іншими словами співвідношення між труднощами завдань мають бути однаковими для будь-якого випробовуваного[2], а співвідношення між рівнями підготовленості випробовуваних – справедливими для завдання будь-якої труднощі. Це означає, що функція (1), залежатиме не від кожного аргументу u і t окремо, а визначатиметься їх відношенням, тобто буде однопараметричною:

$$P = P_1(\zeta),\ \zeta = \frac{u}{t}). \tag{2}$$

[1] За умов виконання правил тестування, що унеможливлюють списування та інші порушення навчальної етики.

[2] Наприклад, тому що для виконання цього завдання потрібно у k раз більше стандартних операцій, ніж для виконання іншого завдання [5].

Щоб з (2) можна було визначити величину ζ, функція $P_I(\zeta)$ має мати взаємозворотну функцію: (3)

$$\zeta = P_1^{-1}(P),$$

яку називають функцією вимірювання, оскільки ймовірність P можна оцінювати за результатами тестування.

Звісно, дійсні види функцій P і P_I – невідомі, і можна говорити тільки про їх наближення деякими моделями. Для цього визначають такі вимоги:

- обидві функції мають бути гладкими та: P – монотонно зростаючою, P_I – монотонно убуваючою;
- $\lim_{\zeta \to 0} P_1(\zeta) = 0$, що означає неможлививість досягнення успіху абсолютно непідготовленим випробовуваним;
- $P = P(u,t)$, $P \in [0,1]$, що гарантує успіх випробуваному, рівень підготовленості якого у багато разів перевищує трудність завдання;
- $P_I(1) = 0{,}5$, тобто максимальна невизначеність у прогнозі результату виконання завдання досягається, коли рівень підготовленості випробовуваного відповідає трудності завдання: $u = t$.

Г. Раш запропонував формулу для функції успіху, яка оцінює імовірність того, що випробовуваний з рівнем підготовленості u правильно виконає завдання з трудністю t [3]:

$$P = P(u \mid t) = \frac{u}{u+t} = \frac{u/t}{1+u/t} = \frac{\zeta}{1+\zeta}. \qquad (4)$$

Відповідно, зворотна їй функція має вигляд т. зв. відношення шансів:

$$\zeta = \frac{P}{1-P} \qquad (5)$$

Трудність j-го завдання (t_j) та рівень підготовленості n-го випробовуваного (u_n) оцінюють за даними матриці відповідей (a_{nj})[4]:

$$t_j = \frac{N - w_j}{w_j} = \frac{1 - \frac{w_j}{N}}{\frac{w_j}{N}} = \frac{1 - p_j}{p_j}; \quad u_n = \frac{Y_n}{M - Y_n} = \frac{\frac{Y_n}{M}}{1 - \frac{Y_n}{M}} = \frac{\pi_n}{1 - \pi_n}, \qquad (6), (7)$$

де w_j – первинний бал j-го завдання (часткова сума матриці відповідей за j-м стовпцем);

Y_n – персональний бал n-го випробовуваного (часткова сума матриці відповідей за n-м рядком), а саме;

$$w_j = \sum_{n=1}^{N} a_{nj}; \quad Y_n = \sum_{j=1}^{M} a_{nj}. \qquad (8), (9)$$

$$a_{nj} = \begin{cases} 1 - n-й\ тестований\ виконав\ j-те\ завдання\ правильно; \\ 0 - n-й\ тестований\ виконав\ j-те\ завдання\ неправильно; \end{cases}$$

M – кількість завдань у тесті;

N – кількість випробовуваних персон;

p_j – частка максимальної кількості балів, отриманих всіма випробовуваними за виконання j-го завдання (оцінка частоти правильного виконання j-го завдання);

π_n – частка максимальної кількості балів, отриманих n-м випробовуваним при виконанні M завдань тесту (оцінка частоти прояви певного рівня підготовленості n-м випробовуваним).

Останні два визначення є інваріантними відносно кількості градацій порядкової шкали первинних оцінок. Наприклад, якщо шкала не дихотомічна, а політомічна[5],

[3] Суворо кажучи, йдеться про статистичне оцінювання умовної (апостеріорної) імовірності рівня підготовленості студента.

[4] Випадкові елементи матриці відповідей a_{nj} є індикаторами успішного рішення n випробовуваним j-го завдання. При дихотомічному оцінюванні умовне маточікування $E(a_{nj}) = P_{nj}$.

[5] У тестів, завдання в яких можуть мати проміжні варіанти відповідей або ж взагалі їх не мати, а відповіді оцінюється в якій-небудь порядковій шкалі.

тобто відповідає набору відповідей, що відрізняються балами, то принцип оцінювання ймовірності отриманого результату через частки набраних балів від можливого залишається незмінним.

Оскільки величини u і t – сумісні, вони повинні належати до однієї шкали та мати загальну одиницю вимірювання, що забезпечує можливість порівняння результатів різних тестів. Г. Раш запропонував використовувати для цього логарифмічну шкалу, тобто перетворення виду:

$$\ln(u) = \theta \; i \; \ln(t) = \beta, \tag{10}$$

що в оберненому вигляді дає:

$$u = \exp(\theta) \;\; t = \exp(\beta). \tag{11}$$

З урахуванням формул (11) функція успіху (4), тобто ймовірність того, що n-й випробуваний надасть правильну відповідь на j-те завдання, набуває вигляду логістичної залежності[6] :

$$P(\vartheta_n, \beta_j) = P_{nj} = \frac{e^{\theta_n}}{e^{\theta_n} + e^{\beta_j}} = \frac{1}{1 + e^{\beta_j - \theta_n}} = \frac{\exp(\theta_n - \beta_j)}{1 + \exp(\theta_n - \beta_j)} = \frac{1}{1 + \exp[-(\theta_n - \beta_j)]}, \tag{12}$$

де θ_n – рівень підготовленості n-го випробовуваного за шкалою Раша;

β_j – рівень трудності j-го завдання за шкалою Раша.

Логістичний характер кривої зберігається й для залежності персонального балу Y_n від значення латентної змінної. Для дихотомічних відповідей, за виконання умови локальної незалежності завдань[7] математичне очікування $E(Y_n)$ первинного балу n-го випробовуваного за моделлю Раша дорівнює:

$$E(Y_n) = \sum_{j=1}^{M} E(a_{nj}) = \sum_{j=1}^{M} P_{nj} = \sum_{j=1}^{M} \frac{1}{1 + \exp(\beta_j - \theta_n)}. \tag{13}$$

За аналогією математичне очікування $E(w_j)$ первинного балу j-го завдання за моделлю Раша дорівнює:

$$E(w_j) = \sum_{n=1}^{N} E(a_{nj}) = \sum_{n=1}^{N} P_{nj} = \sum_{n=1}^{N} \frac{1}{1 + \exp(\beta_j - \theta_n)}. \tag{14}$$

Діленням суми (13) на кількість завдань, отримують середню ймовірність, або очікуваний відсоток правильних відповідей в залежності від значення латентної змінної θ.

У вигляді (12) функція успіху має один параметр – різницю θ_i - β_j і властивості, що відповідають вимогам моделі.

З урахуванням формули (5) рішенням рівняння (12) буде:

$$E(w_j) = \sum_{n=1}^{N} E(a_{nj}) = \sum_{n=1}^{N} P_{nj} = \sum_{n=1}^{N} \frac{1}{1 + \exp(\beta_j - \theta_n)}. \tag{15}$$

яке можна обчислити за умов $P_{nj} \neq 0$ і $P_{nj} \neq 1$.

Не важко показати, що різницю між рівнями підготовленості n-го та l-го випробовуваних, що відповідали на одне й теж саме j-те питання, можна оцінити у такий спосіб:

$$\theta_n - \theta_l = \ln\left(\frac{P_{nj}}{1 - P_{nj}}\right) - \ln\left(\frac{P_{lj}}{1 - P_{lj}}\right). \tag{16}$$

[6] Інша назва – основна логістична модель Раша (вона покладена в основу т. зв. *IPL*-моделі *IRT*, або "1 Parametric Logistic Latent Trait Model") [7].

[7] Під локальною (умовною) незалежністю завдань розуміють те, що зв'язок між ними можливий виключно опосередковано через латентну змінну. Це припущення *IRT* разом з припущенням щодо монотонності функції $P(g, t)$ є настільки важливим, що його розглядають в якості критерію для визначення розмірності первинних даних – мінімальної кількості латентних ознак, необхідних для пояснення відносин між завданнями та випробовуваними.

Очевидно, що рівняння (16), якщо дані відповідають обраній моделі, не повинно залежати від j, тобто від вмісту завдання. Ця умова має назву "item-free". Аналогічне міркування приводить до незалежності різниць ("відстаней") між труднощами завдань і рівнями підготовленості випробовуваного ("person-free").

Раш обмежив суму параметрів трудності всіх завдань нулем (умова параметризації, або калібрування завдань):

$$\sum_{j=1}^{M} \beta_j = 0, \tag{17}$$

встановивши тим самим масштаб для параметра θ.

Формули (15) і (17) представляють процедуру шкалювання за Рашем, тобто спосіб трансформації результатів тестування (6) і (7) з порядкової в інтервальну шкалу (т. зв. RM – шкалу)[8] . Одиниця RM-шкали отримала назву "логіт". Застосування логітів у IRT-моделях як міри підготовленості випробовуваних і міри трудності завдань дає ряд переваг. Найважливішою є та, що результат вимірювання в шкалі Раша значно менше залежить від вибірки випробовуваних (person free measurement) і набору питань тесту (item free calibration) ніж у СТТ. Крім того, логіт – метрична одиниця, яка забезпечує об'єктивне порівняння за єдиною шкалою досягнень різних студентів з різних учбових дисциплін, надає можливість підсумовувати ці досягнення та будувати об'єктивні рейтинги [6; 7].

Реальний діапазон логіт-оцінок невеликий: від -5 до +5. Середнє арифметичне логіт-оцінок дорівнює нулю. Його приймають за початок відліку. Нульова оцінка відповідає ситуації, коли 50% випробовуваних виконують правильно завдання, а 50% – неправильно. При зменшенні трудності завдання на 1 логіт імовірність правильної відповіді збільшується приблизно на 25%, а при збільшенні – зменшується приблизно на 25%. Таким чином, 1 логіт приблизно визначає квантиль рівня Q_{nj}. Значення -2,94 логіта відповідає простому завданню, яке 95% випробовуваних виконують правильно, а 5% – неправильно. Значення +2,94 логіта, навпаки, відповідає важкому завданню, що під силу 5% випробовуваних. Логіт-оцінки часто розрізняються у другому та третьому знаку після коми і малопридатні для повідомлення випробовуваним. Тобто ситуація близька до такої з Z-шкалою. Тому шкалу логітів подібно до Z-шкали піддають лінійним перетворенням для приведення оцінок до цілого невід'ємного вигляду. Наприклад, перетворення виду:

$$\widetilde{\theta}_n = Y_M \cdot \exp(\theta_n)/[1+\exp(\theta_n)], \tag{18}$$

де Y_M – максимальний бал тесту;
переносить значення θ_n у діапазон від 0 до Y_M.

За достатньої чисельності випробовуваних є сенс говорити про відносні частоти (як оцінки відповідних імовірностей) вірних рішень ($\hat{P}_j(y)$) та невдач ($\hat{Q}_j(y)$) при виконанні j-го завдання випробовуваними, що набрали однаковий індивідуальний бал y y ($0 \le y \le Y_M$):

$$\hat{P}_j(y) = \hat{P}_{yj} = N_{yj}/N_y \text{ і } \hat{Q}_j(y) = 1 - \hat{P}_j(y) = (N_y - N_{yj})/N_y, \tag{19}$$

де N_{yj} – кількість випробовуваних, що виконали j-те завдання й отримали однаковий персональний бал y;

N_y – кількість випробовуваних, що отримали однаковий персональний бал y.

Оцінка дисперсії величини $\hat{P}_j(y)$ з формули (19) дорівнює:

$$s^2[\hat{P}_j(y)] = \frac{\hat{P}_j(y)\cdot\hat{Q}_j(y)}{N_y - 1} = \frac{N_{yj}\cdot(N_y - N_{yj})}{N_y^2\cdot(N_y - 1)}. \tag{20}$$

Оскільки за моделлю Раша рівень підготовленості всіх N_y випробовуваних однаковий $\theta_{N_y} = \hat{\theta}_y$, у термінах (19) функція успіху вирішити j-те завдання випробовуваним, який набрав y балів при дихотомічному оцінюванні, має вигляд:

[8] Метрична система вимірювань Раша.

$$\hat{\theta}_y - \hat{\beta}_j(y) = \hat{\zeta}_j(y) = \ln\frac{\hat{P}_j(y)}{\hat{Q}_j(y)};\ j\in\{1, 2,\ldots, M\};\ y\in\{0, 1, \ldots, M\}, \quad (21)$$

а ймовірність успіху:

$$P_j(y) = P_{yj} = \frac{\exp(\hat{\theta}_y - \hat{\beta}_j)}{1+\exp(\hat{\theta}_y - \hat{\beta}_j)} = \frac{1}{1+\exp[-(\hat{\theta}_y - \hat{\beta}_j)]}. \quad (22)$$

Як і у формулі (16), у формулі (21) для забезпечення логарифмування має бути $\hat{P}_j(y) \neq 0$ і $\hat{P}_j(y) \neq 1$. Тому для отримання рішення за формулами (16) чи (20) до кожного екстремального первинного балу $y = 0$ або $y = M$, звичайно додають невелику константу, наприклад, 1/150 [4].

Оцінка дисперсії величини $\hat{\theta} - \hat{\beta}_j(y)$, як показано у роботі [4], дорівнює:

$$s^2[\hat{\theta} - \hat{\beta}_j(y)] = s^2[\hat{\zeta}_j(y)] = \frac{(N_y+1)\cdot(N_y+2)}{N_y\cdot(N_{jy}+1)\cdot(N_y - N_{jy}+1)}. \quad (23)$$

Якщо знехтувати кореляцією між результатами виконання завдань, то формула (23) набуває вигляду [4]:

$$s^2(\hat{\theta} - \hat{\beta}_j) = s^2(\hat{\zeta}_j) = \frac{1}{M^2}\cdot\sum_{y=0}^{M} s^2[\hat{\zeta}_j(y)] = \frac{\sum_{y=0}^{M} W_y \cdot s^2[\hat{\zeta}_j(y)]}{\sum_{y=0}^{M} W_y} = s^2(\hat{\beta}_j), \quad (24)$$

де M – максимальний персональний бал тесту;

W_y – статистична вага набраного бала у, тобто частота появи у серед всіх балів тесту ($0 \le W_y \le 1$).

$$s^2(\hat{\theta} - \hat{\beta}) = s^2(\hat{\zeta}) = \frac{1}{M^2}\cdot\sum_{j=1}^{M} s^2(\hat{\zeta}_j), \quad (25)$$

і

$$s^2[\hat{\theta}(y)] = s^2[\hat{\zeta}(y)] + s^2(\hat{\zeta}_j). \quad (26)$$

Імовірність P_{nj} як функція β_j за фіксованого рівня підготовленості $\theta_n = \theta_0$ повністю описує (прогнозує) потенціальні можливості n-го випробовуваного з рівнем підготовленості θ_0 при виконанні завдань всіх можливих труднощів β_j, тому її називають характеристичною функцією рівня підготовленості θ_0 (рис. 1, $P(\theta_0$=-3, $\beta)$; $P(\theta_0$=1, $\beta)$; $P(\theta_0$=3, $\beta)$.

Так само ймовірність P_{nj} як функція θ_n за фіксованого значення β_j характеризує можливості випробовуваних з різним рівнем підготовленості при виконанні завдання труднощі $\beta = \beta_0$ і називається характеристичною функцією труднощі β_0 (рис. 1), $P(\theta$, β_0=-3); $P(\theta, \beta_0$=1); $P(\theta, \beta_0$=3) [4]. Абсциси $\beta = \theta_0$ і $\theta = \beta_0$ на рис. 1 є розв'язками рівнянь: $\frac{\partial^2 P}{\partial\beta^2} = 0$ і $\frac{\partial^2 P}{\partial\theta^2} = 0$ для *1PL* моделі, тобто представляють точки перетину характеристичних функцій.

З рис. 1 бачимо, що зміна величини θ_0 приводить до зміщення кривої $P(\beta, \theta_0)$ без деформації паралельно вісі абсцис на θ_0 від нуля. Ординати точок перегину завжди дорівнюють 0,5. Це значення зберігається для всіх величин θ_0, тому графіки $P(\beta, \theta_n)$ результатів тестування n-го випробовуваного за завданнями різного рівня трудності повинні мати вигляд неперехресних кривих однієї форми. Аналогічна ситуація спостерігається з кривими $P(\beta_j, \theta)$, тобто завдання різного рівня труднощі мають створювати родину неперехресних S-образних кривих.

Чи відбувається це в дійсності, залежить від узгодженості відповідей вимогам *1PL*-моделі. А саме від того, чи є вони локально незалежними одна від іншої, чи мають завдання однакові роздільні здатності й чи дійсно вони вимірюють одну й ту ж саму

латентну рису випробовуваних. Крім того, щоб отримати стійкі результати, кількість випробовуваних, що мають однакову оцінку латентної характеристики, повинна бути значною. Останнє вважають за недолік *1PL*-моделі, а також *IRT* в цілому.

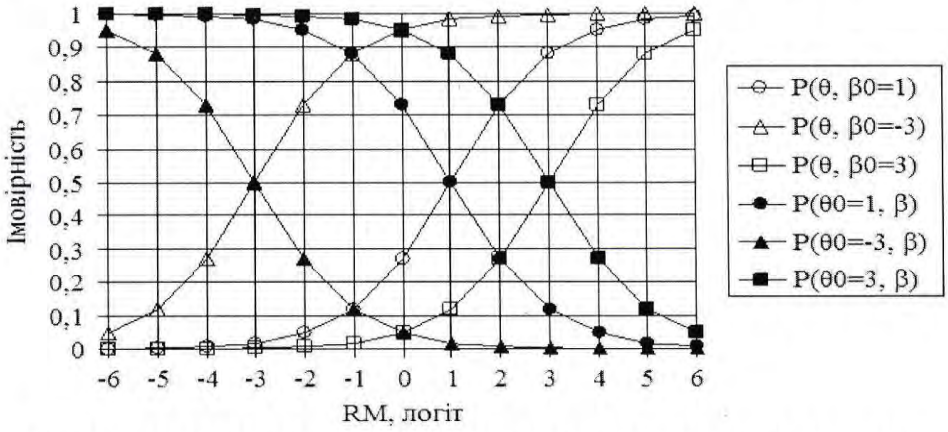

Рисунок 1. *IRT*-**характеристики різних рівнів підготовленості** θ_0 **(світлі маркери) та рівнів труднощі** β_0 **(темні маркери).**

Джерело: розраховано і побудовано автором

Графіки $P(\beta,\theta_0)$, $P(\theta,\beta_0)$ дуже слушними для оцінювання властивостей тесту і тому мають назву характеристичних кривих (Item characteristic curve – ICC) відповідно: латентної змінної, завдання та тесту в цілому. За Ф. Лордом [8] ICC для тесту в цілому представляє регресію персонального балу Y на латентну змінну θ[9].

Очевидно, що чим більш крутою є характеристична крива завдання, тим більш вузьким є інтервал, на якому це завдання диференціює випробовуваних за рівнем їх підготовленості, й тим більш високою є роздільна (диференціювальна, дискримінаційна) здатність завдання.

До речі, окрім статистичних методів визначення роздільної здатності тесту, про які йшла мова раніше [2], якщо підходити до цього поняття як до довжини проміжку $\Delta\theta$ у логітах на шкалі рівня підготовленості, який відповідає кроку ΔY=1[10], то роздільну здатність можна оцінити *apriori*. Для цього достатньо обчислити зворотне значення часткової похідної функції $Y_n = f(\theta_n)$ по θ_n і прийняти $\partial Y_n = 1$. Наприклад, для моделі Раша:

$$\Delta\theta_n \approx \frac{\partial\theta_n}{\partial Y_n} = 1/\left(\sum_{j=1}^{M} P_{nj}\cdot Q_{nj}\right), \tag{27}$$

що приводить до оцінок $4/M < \Delta\theta < 11/M$ з СКВ $s(\Delta\theta) \approx 2/M$.

А. Бірнбаум [9] увів у модель Раша другий латентний параметр (d_j)[11]:

$$P_{nj} = \frac{\exp[d_j\cdot(\theta_n-\beta_j)]}{1+\exp[d_j\cdot(\theta_n-\beta_j)]} = \frac{1}{1+\exp[-d_j\cdot(\theta_n-\beta_j)]}. \tag{28}$$

Цей параметр отримав назву критерію роздільної здатності завдання, або нахилу ICC завдання. В залежності від величини d_j ICC завдання може змінювати форму від прямої до стрибкоподібної (рис. 2).

Часто в обчислювальних цілях до умови (17) для *2PL*-моделі додається умова параметризації:

$$\prod_{j=1}^{M} d_j = 1. \tag{29}$$

[9] Побудова множинної лінійної регресії Y на бінарну змінну a_{nj} є некоректним, оскільки лінійна регресія припускає недихотомічність a_{nj}. Логістичне перетворення знімає цю проблему.

[10] Тобто якщо $|\theta_1 - \theta_2| \le \Delta\theta$, то розрізнити θ_1 і θ_2 засобами даного тесту неможливо.

[11] Прийнята назва – двопараметрична модель Бірнбаума (або *2PL IRT-model*).

Двопараметрична *IRT*-модель (28) дозволяє за результатами пробного тестування відбирати завдання, оптимальні як за трудністю (за положенням на вісі абсцис), так і за роздільною здатністю (за нахилом кривої). Вироджена у пряму ICC (поз. 2 на рис. 2) явно не підходить через недостатню роздільну здатність, а стрибкоподібна (поз. 3 на рис. 2) – дуже ефективно диференціює тих, у кого оцінки менше 1,45, та тих, хто отримав більше 1,55 логіта, але це занадто вузький діапазон. Оптимальним можна вважати завдання з роздільною здатністю 0,6–2,0, оскільки вони охоплюють оцінки від -5 до +5 логітів.

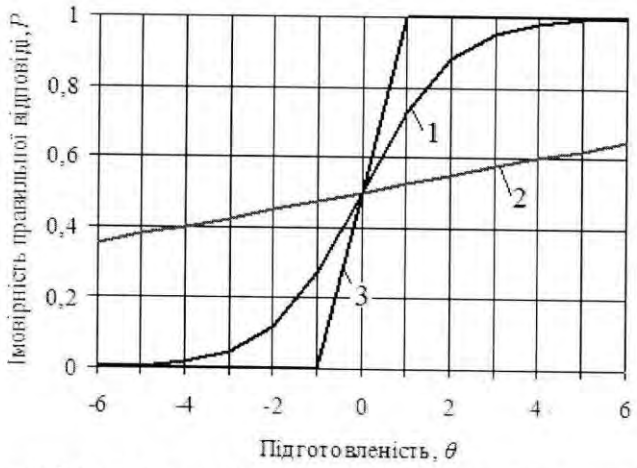

Рисунок 2. Характеристичні криві трудності завдань $\theta_n = \beta_0$ з різними роздільними здатностями: 1) $d_j = 1$; 2) $d_j = 0{,}1$; 3) $d_j = 100$.

Джерело: розраховано і побудовано автором

Часто вважають, що параметр нахилу лінійно пов'язаний (за деяких умов) зі змінною навантаження у факторному аналізі. Як відмічають автори [10], за припущенням нормального розподілу θ:

$$d_j \cong \frac{\rho_j}{\sqrt{1-\rho_j^2}}, \tag{30}$$

де ρ_j – коефіцієнт кореляції відповідей на паралельні завдання.

Бок і Айткін [11] звернули увагу на те, що вираз (28) можна представити у вигляді:

$$P_{nj} = \{1+\exp[-(c_j + d_j \cdot \vartheta_n)]\}^{-1}, \tag{31}$$

де $c_j = - d\beta_j$.

Відповідно для моделі *1PL* маємо:

$$P_{nj} = \{1+\exp[-(c_j + d \cdot \vartheta_n)]\}^{-1}, \tag{32}$$

де $c_j = - d.\beta_j$.

Логарифмуванням, наприклад, виразу (32) отримуємо рівняння:

$$\lg\left(\frac{P_{nj}}{1-P_{nj}}\right) = c_j + d \cdot \theta_n, \tag{33}$$

що представляє рівняння регресії зі змішаними впливами (багаторівнева, або ієрархічна модель: відповіді на питання розглядають як рівень1, що вкладений у рівень 2 завдань). Для оцінювання параметрів цього рівняння можна скористатись статистичними пакетами: *SAS (PROC NLMIXED), USEPA, Stata* тощо.

Як відомо, логістичні криві відносяться до родини функцій, що мають розподіл виду:

$$F(Y) = (1+e^{-dY})^{-1} \tag{34}$$

з дисперсією:

$$\sigma^2(Y) = \frac{1}{3}\left(\frac{\pi}{d}\right)^2 \qquad (35)$$

де d – параметр, що визначає нахил (похідну) кривої в точці перегину.

Важливо, що за $d = 1{,}7$ функція (28) є майже тотожною нормальній стандартній інтегральній кривій розподілу ймовірності (нормальній огіві) N(0, 1)). В цьому разі, якщо бали випробовуваного надані у логітах чи Z-оцінках, площа під огівою буде завжди менше 1, й огіву можна надійніше трактувати як частку випробовуваних, які успішно впоралися із завданням, в залежності від рівня θ.

Тому модель *2PL* в основному застосовують при значенні a, близькому до 1,7. Зокрема, у (28) замість d використовують $1{,}7d$, що дозволяє оперувати з моделями *1PL* і *2PL* при $0{,}9 \le d \le 1{,}1$ з позицій добре вивченого нормального закону розподілу ймовірності [4].

Згодом А. Бірнбаум удосконалив *2PL*-модель для завдань закритої форми, увівши до неї з метою компенсації ймовірності вгадування правильної відповіді третій параметр c_j[12]:

$$P_{nj} = c_j + (1 - c_j)\cdot\{1 + \exp[-d_j\cdot(\theta_n - \beta_j)]\}^{-1}, \qquad (36)$$

де c_j – імовірність вгадування правильної відповіді на j-е завдання[13].

Очевидно, що межа функції (36) за $\theta_n \to -\infty$ дорівнює c_j, тобто її графік має горизонтальну асимптоту $P(-\infty) = c$. Завдяки цьому початок *ICC* у *3PL*-моделі виявляється зміщеним вгору уздовж вісі ординат на величину c_j, що визиває її виродження, і відповідно, зменшення роздільної здатності (рис. 3).

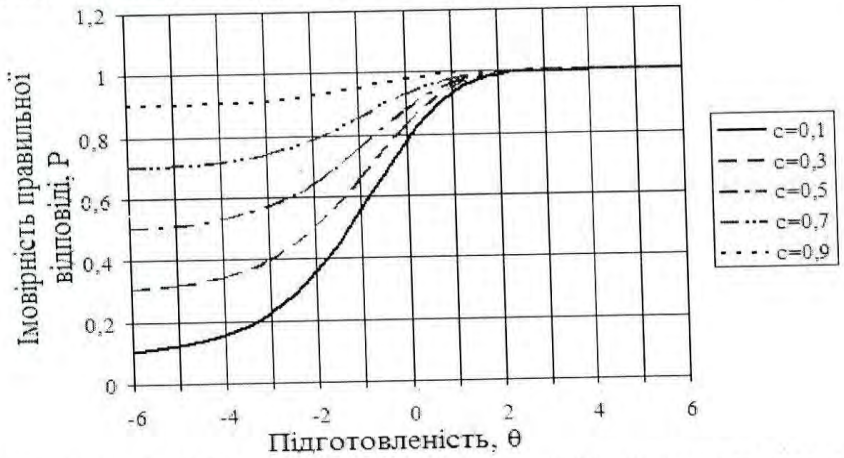

Рисунок 3. ICC з різними значеннями параметру вгадування.

Джерело: розраховано і побудовано автором

Якщо $\beta = \theta$, то випробувані мають $(0{,}5 + 0{,}5c)\cdot 100\%$ шансів правильно вирішити завдання [7]. Параметру c зазвичай задають початкове значення $1/K$, де K – кількість відповідей на завдання.

3PL-модель складна для реалізації (*ICC* часто перехрещуються, тому до підбору завдань треба підходити ретельніше) і має меншу точність. Мабуть, тому *2PL*-модель набула більшої популярності у навчальному тестуванні.

Одним з найважливіших показників якості *IRT*-моделей є інформаційна функція Фішера, що визначає кількість інформації $I(\theta_n)$ про латентну змінну θ, що міститься у первинному балі n-го випробовуваного [12]:

$$I(\vartheta_n) = \sum_{j=1}^{M} I(\vartheta_n, \beta_j), \qquad (37)$$

[12] Трипараметрична модель Бірнбаума (або *3PL IRT-model*).

[13] Іноді c_j називають також рівнем псевдо-успіху, або ймовірністю правильно виконати завдання для тих, хто має мінімальний рівень підготовленості.

де

- для *1PL*-моделі:

$$I_j(\vartheta_n, \beta_j) = P(\vartheta_n, \beta_j) \cdot Q(\vartheta_n, \beta_j), \tag{38}$$

- для *2PL*-моделі:

$$I_j(\vartheta_n, \beta_j) = d_j^2 \cdot P(\vartheta_n, \beta_j) \cdot Q(\vartheta_n, \beta_j), \tag{39}$$

- для *3PL*-моделі:

$$I_j(\vartheta_n, \beta_j) = d_j^2 \cdot \frac{[(P(\vartheta_n, \beta_j) - c_j]^2}{(1-c_j)^2} \cdot \frac{Q(\vartheta_n, \beta_j)}{P(\vartheta_n, \beta_j)}; \tag{40}$$

$I_j(\vartheta_n, \beta_j)$ – кількість інформації щодо різниці $\theta_n - \beta_j$, що міститься у a_{nj}-му елементі матриці відповідей.

Іншими словами, інформаційна функція характеризує точність вимірювання для осіб, які перебувають на різних рівнях θ (рис. 4); більш висока інформаційність позначає більш високу точність, тобто визначає тих, для кого конкретне завдання (рис. 4а) або тест у цілому (рис. 4б) є найкориснішим з точки зору диференціації рівня підготовленості.

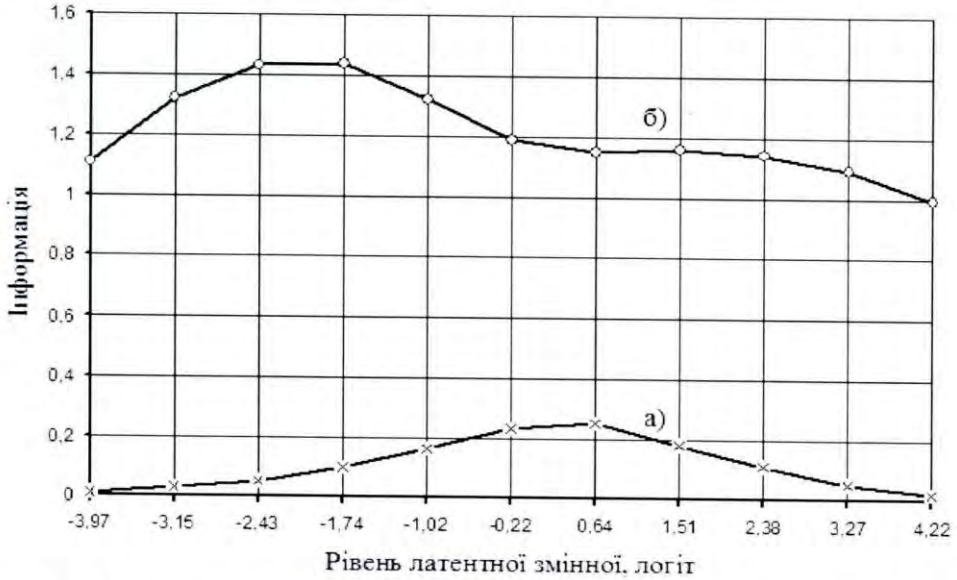

Рисунок 4. Приклад інформаційної функції завдання (а) та тесту в цілому (б).

Джерело: розраховано і побудовано автором

Форма інформаційної функції завдання залежить від його параметрів. Чим вище дискримінаційна можливість завдання, тим більш чітко виражений пік матиме інформаційна функція; тобто збільшення параметра дискримінації підвищує точність інформації щодо випробовуваних, чиї рівні латентної характеристики θ лежать поруч зі значенням β трудності завдання.

На кожному рівні θ інформаційна функція приблизно дорівнює очікуваному значенню оберненого квадрату стандартної помилки θ-оцінки [8]. Чим менше стандартна помилка вимірювання (*SE*):

$$s(\hat{\theta}) = SE(\hat{\vartheta}) = \frac{1}{\sqrt{I(\hat{\theta})}}, \tag{41}$$

тим більше інформації або більш високу точність шкали забезпечує оцінка θ (див. рис. 4). Очевидно, що найвища точність досягається у середині *RM*-шкали (для "хорошистів"), а θ-оцінки на краях шкали стають нескінченно великими. Тобто найважче розрізняти "неуків" та "відмінників".

Важливо, що на відміну від *CTT*, де однакові рейтинги представляють однаковий рівень латентної риси θ, в *IRT* береться до уваги, які саме завдання були виконані правильно, а які – помилково. Особи з однаковими сумарними балами, але з різними зразками відповідей можуть мати різні *IRT*-оцінки θ. Людина може відповісти на більш дискримінаційні та складні завдання і отримати більш високу оцінку θ, ніж та, яка відповіла на таку ж кількість завдань з низькою дискримінацією або складністю.

Описані вище моделі можна вважати фундаментом Раш-оцінювання. На їх основі на сьогодні розроблено більш складні й точні математичні моделі тестування (детальний огляд див., напр., [13]). Серед найвідоміших Раш-моделей, що використовують політомічний формат відповіді, відмітимо:

1. *Graded Model* (модель з фіксованими категоріями (рівнями) відповідей, тобто коли відповідь прив'язана к певному балу; автор F. Samejima).

Ця модель стала базовою для інших моделей, оскільки в ній вперше було зафіксовано поняття рівня у визначенні ймовірності того, що n-й випробовуваний схвалить k-ту відповідь на j-те завдання:

$$P(a_{nj}=k\,|\,\theta_n)=\{1+\exp[-\mathrm{d_j}\cdot(\theta_\mathrm{n}-\beta_{\mathrm{jk}})]\}^{-1}-\{1+\exp[-\mathrm{d_j}\cdot(\theta_\mathrm{n}-\beta_{\mathrm{jk+1}})]\}^{-1}, \qquad (42)$$

де $k = 1,\ldots, K_j$;

K_j – кількість відповідей у j-му завданні;

d_j – параметр дискримінації j-го завдання;

β_{jk} – параметр труднощі j-го завдання на k-му кроці його виконання (k-й поріг), при цьому $\beta_{j1} < \ldots < \beta_{jk} < \ldots \beta_{jKj-1}$ і $\beta_{jKj} = \infty$

2. *Nominal Model* (номінальна модель; розвиває попередню модель і дозволяє визначити, які альтернативні відповіді пов'язані з більш високим рівнем θ; автор R. D. Bock).

3. *Partial Credit Model* (модель часткового оцінювання, або модель оцінювання завдань, що складаються з декількох питань, розв'язуваних послідовно, оцінки за які залежить від кількості правильних відповідей (подоланих порогових кроків); автор G. N. Masters [14]).

4. *Rating Scale Model* (модель рейтингової шкали; всі завдання мають однакову кількість категорій оцінювання, трудність і роздільну здатність; автор D. Andrich).

5. *Many- facet Rasch Model* (багатопараметрична *IRT*-модель; представляє розширення моделі Раша з довільними проміжними категоріями виконання завдань на випадок, коли до розрахунку включаються не тільки параметри завдань та випробовуваних, а й параметри експертів, які оцінюють виконання завдань; автор J. M. Linacre).

6. *Multidimensional Rasch Model* (багатовимірна модель, в якій рівень підготовки випробовуваного визначається як лінійна комбінація або добуток декількох прихованих змінних; автори M. Wilson та інші. Детальний огляд наведено у [13; 14].

Крім того, існує клас *IRT*-моделей, в яких припущення монотонності та локальної незалежності завдань послаблені або зняті. Це моделі, що розкладаються або містять обмеження на параметри завдань, клас непараметричних моделей та інш. [13].

Коротко зупинимось на моделі часткового оцінювання (*PCM*) [14], що являє собою адаптацію *2PL*-моделі для політомічних тестів.

Згідно з *PCM*-моделлю умовна ймовірність того, що n-й випробуваний при виконанні j-го завдання правильно відповість на g питань (виконає g кроків) і набере g балів, дорівнює:

$$P(Y_j=g\,|\,\theta_n, \textit{параметри завдань})=P_{njg}=\frac{\exp[\sum_{h=0}^{g}1{,}7\cdot d_j\cdot(\theta_n-\delta_{jh})]}{\sum_{l=0}^{m_j}\exp[\sum_{h=0}^{l_j}1{,}7\cdot d_j\cdot(\theta_n-\delta_{jh})]}=$$

$$=\frac{\exp\{1{,}7\cdot d_j\cdot[g\cdot\theta_n-\sum_{h=0}^{g}\delta_{jh}]\}}{\sum_{l_j=0}^{m_j}\exp\{1{,}7\cdot d_j\cdot[l_j\cdot\theta_n-\sum_{h=0}^{l_j}\delta_{jh}]\}}, \qquad (43)$$

де d_j – параметр роздільної здатності j-го завдання;
m_j – кількість варіантів (категорій) відповідей у j-му завданні;
δ_{jh} – рівень труднощі (поріг) h-го кроку виконання j-го завдання[14];
l_j – максимально можлива оцінка за виконання j-го завдання; $l_j = m_j$;
g – кількість балів, набраних за виконання j-го завдання; $g = 0, 1, \ldots, l_j$.
Для спрощення розрахунків приймають:

$$\delta_{j0} \equiv \theta_n \text{ і } \exp[\sum_{h=0}^{g}(\theta_n - \delta_{jh})] = \exp[\sum_{h=1}^{g}(\theta_n - \delta_{jh})]. \quad (44)$$

Не важко показати, що:

$$\ln(\frac{P_{njg}}{P_{njg+1}}) = \ln\{\exp[\sum_{h=0}^{g}1{,}7 \cdot d_j \cdot (\theta_n - \delta_{jh})]\} - \ln\{\exp[\sum_{h=0}^{g}1{,}7 \cdot d_j \cdot (\theta_n - \delta_{jh})]\} = \quad (45)$$
$$= 1{,}7 \cdot d_j \cdot (\theta_n - \delta_{jg}),$$

Формула (45) представляє компактний запис *PCM*-моделі, придатний для лінійного оцінювання.

На рис. 5 наведено приклад *ICC* для завдання з m_j=4, δ_{jh}={-3; -0,5; 0} і d_j=1/1,7. Легко бачити, що перетини *ICC*, які представляють імовірності отримання оцінок g і g+1, відбуваються у точках $\theta_n = \delta_{jh}$, абсциси яких відповідають трудності виконання відповідного кроку. Наприклад, студент із рівнем підготовленості θ = -3 з рівною ймовірністю (P=0,5) отримає за завдання оцінку 0 або 1, тобто поріг трудності переходу з категорії оцінок 0 у категорію оцінок 1 сягає -3 логіта.

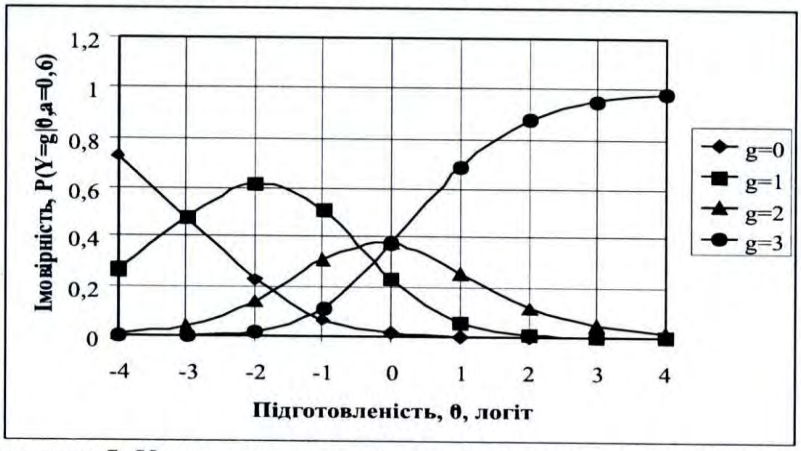

Рисунок 5. Характеристичні криві трикрокового завдання.
Джерело: розраховано і побудовано автором

За припущенням локальної незалежності завдань ("чистоти тесту") та наявності однієї (базової) латентної змінної, що визначає ймовірність правильної відповіді випробуваного, спільна умовна ймовірність отримання Y балів у тесті з M завдань, або функція правдоподібності дискретної величини Y, дорівнює:(46)

$$L(Y \mid \theta_n, \textit{параметри завдань}) = \prod_{j=1}^{M}\prod_{g=0}^{m_j} P_{njg}^{u_{jg}},$$

де u_{jg} – індикаторна змінна, що визначається, як:

$$u_{jg} = \begin{cases} 1 & \textit{якщо } Y_j \textit{ дорівнює } g; \\ 0 & \textit{у протилежному випадку.} \end{cases}$$

Знаходження максимуму функції (46) шляхом підбору параметрів моделі є чи не найпопулярнішим алгоритмом, що отримав назву методу максимальної

[14] У дихотомічному варіанті оцінювання завдань існує тільки дві категорії оцінок: 0 та 1. Відповідно $m_j = 2$; $\delta_{jh} = d_j$.

правдоподібності (*likelihood*). В обчислювальній математиці задача визначення параметрів моделі процесу за вимірюваннями його виходу називається зворотною і належить до класу некоректно поставлених. Основною її проблемою є сильна чутливість розв'язку до невеликих відхилень вхідних даних (див., напр., [15]), тому особлива увага приділяється стабілізації розв'язку.

У роботі [16] стверджується, що *PCM*-модель придатна для оброблення відповідей на завдання на встановлення відповідності з можливістю вибору декількох правильних відповідей та в інших практичних випадках. Однак слід відмітити, що можливості застосування *PCM*-моделі (42) до результатів реальних педагогічних тестів багатьма дослідниками ставились під сумнів (див. напр., огляд [17]). Зокрема, вони відмічали такі її недоліки: персональні бали у *PCM* лінійно залежать від індексу категорії, а всі *ICC* повинні мати однаковий нахил. Згодом останнє обмеження було знято у т. зв. узагальненій *PCM*-моделі, або *GPCM*, яку запропонував Е. Муракі [18]:

$$P(y_j = g \mid \theta_n, \textit{параметри завдань}) = P_{njg} = \frac{\exp[\sum_{h=0}^{g} 1{,}7 \cdot d_j \cdot (\theta_n - \beta_j - \delta_{jh})]}{\sum_{l_j=0}^{m_j} \exp[\sum_{h=0}^{l_j} 1{,}7 \cdot d_j \cdot (\theta_n - \beta_j - \delta_{jh})]} \quad (47)$$

де β_j – як і у попередніх формулах, показник трудності j-го завдання;
δ_{jh} – значення порогу трудності категорії g в j-му завданні, а інші позначення співпадають з такими у формулі (42). При цьому для подалання невизначеності встановлюють умови параметризації:

$$\delta_{j0} = 0 \text{ і } \sum_{h=1}^{m_j} \delta_{jh} = 0. \quad (48)$$

На моделі *PCM* і *GPCM* дуже схожа модель рейтингової шкали. Її відмінність полягає в тому, що крок за *RM*-шкалою між рівнями трудності категорій в ній для кожного завдання є однаковим:

$$P(y_j = g \mid \theta_n, \textit{параметри завдань}) = P_{njg} = \frac{\exp[\sum_{h=0}^{g} (\theta_n - \beta_j - b_h)]}{\sum_{l=0}^{M} \exp[\sum_{h=0}^{M} (\theta_n - \beta_j - b_h)]} =$$

$$= \frac{\exp[g \cdot (\theta_n - \beta_j) - \sum_{h=0}^{g} b_h)]}{\sum_{l=0}^{M} \exp[l \cdot (\theta_n - \beta_j) - \sum_{h=0}^{M} b_h)]}, \quad \textit{або} \quad \ln(\frac{P_{njg}}{P_{njg+1}}) = \theta_n - \beta_j - b_g, \quad (49)$$

де $$\sum_{h=0}^{0} (\theta_n - b_j - \beta_h) = 0;$$

b_j – рівень (поріг) труднощі j-го завдання за *RM*-шкалою.

Порівнюючи (42) і (49), наприклад, бачимо, що перехід між ними здійснюється замінами:

$$g \cdot \theta_n - \sum_{h=0}^{g} \delta_{jh} \Rightarrow g \cdot (\theta_n - \beta_j) - \sum_{h=0}^{g} b_h \quad l \cdot \theta_n - \sum_{h=0}^{g} \delta_{jh} \Rightarrow l \cdot (\theta_n - \beta_j) - \sum_{h=0}^{g} b_h.$$

GPCM-модель сьогодні є одним з основних інструментів оцінювання досягнень студентів і випускників навчальних закладів службами тестування США (NAEP, SAT, GRE), а також неодноразово використовувалась у великомасштабних дослідженнях, у тому числі *у Національному оцінюванні навчального процесу (NAEP) у США, Національному обстеженні рівня письменності дорослих (NALS) у США, міжнародних оглядах письменності та життєвих навичок дорослого населення (IALS/IALLS)*. З використанням *IRT*-моделей працюють багато організацій з сертифікації спеціалістів з програмного та апаратного забезпечення, таких *як European Computer Driving License Foundation або Thomson Prometric.*

(продовження слідує)

Список використаних джерел

1. Сініцький М. Є. Статистичні інструменти вимірювання якості освіти. Ч. 1. Класичний підхід // Науковий вісник НАСОА. – 2014. – № 4. – С. 58–69.
2. Сініцький М. Є. Статистичні інструменти вимірювання якості освіти. Ч. 2. Класичний підхід // Науковий вісник НАСОА. – 2015. – № – С.
3. Rash G. On Objectivity and Specificity of the Probabilistic Basis for Testing [Electronic resource]. – Access mode : http://www.rasch.org/memo196x.pdf.
4. Нейман Ю. М. Введение в теорию моделирования и параметризации педагогических тестов / Ю. М. Нейман, В. А. Хлебников. – М. : Прометей, 2000. – 168 с.
5. Крокер Л. Введение в классическую и современную теорию тестов : учебник / Л. Крокер, Дж. Алгина ; пер. с англ. Н. Н. Найденовой, В. Н. Смилкина, М. Б. Челышковой ; под общ. ред. В. И. Звонникова, М. Б. Челышковой. – М. : Логос, 2010. – 668 с.
6. Аванесов В. С. Item Response Theory: Основные понятия и положения [Электронный ресурс]. – Режим доступа : testolog@mail.ru.
7. Bryce B. Reeve. An Introduction to Modern Measurement Theory. – Outcomes Research Branch Applied Research Program, Division of Cancer Control and Population Sciences [Electronic resource]. – Access mode : http://www.applied¬research.cancer.gov/archive/.../immt.pdf.
8. Lord F. M. Application of Item Response Theory to Practical Testing Problems / F. M. Lord. – Hillsale, NJ : Lawrence Erlbaun Associates, 1980.
9. Birnbanm A. Some Latent Trait Models and Their Use in Imferring an Examinee's Ability / A. Birnbahn // F. M. Lord, M. P. Novick. Statistical Theories of Mental Test Scores. – Reading MA : Addison-Wesley, 1968.
10. Linden W. Handbook of Modern Item Response Theory / W. Linden, R. Habbleton. – NY. : Springer-Verlag, 1997.
11. Bock R. D. Marginal Maximum Likelihood Estimation of Item Parameters: An Application of the em Algorithm / R. D. Bock, M. Aitkin // Psychometrika. – 1981. – No 46. – P. 443–459.
12. Item Response Theory [Electronic resource]. – Access mode : www://en.wikipedia.org/wiki/Item_response_theory#Information.
13. Sijtsma K. Item Response Theory: Past Performance, Present Developments, and Future Expectations / K. Sijtsma, B. W. Junker // Behaviometrika. – 2006. – Vol. 33, No 1. – P. 75–102.
14. Masters G. N. A Rasch Model for Partial Credit Scoring / G. N. Masters // Psychometrika. – 1982. – Vol. 47, No 1. – P. 149–174.
15. Bartholomew D. The Analysis and Interpretation of Multivariate Data for Social Scientists / D. Bartholomew, F. Steele, I. Moustaki, J. Galbraith. – Charman & Hall : London, 2002.
16. Skrondal A. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models / A. Skrondal, S. Rabe-Hesketh. – Charman & Hall : Boca Ration, FL, 2004.
17. Лоусон Ч. Численное решение задач наименьших квадратов / Ч. Лоусон, Р. Хенсон ; пер. с англ. – М. : Наука. Гл. ред. физ.-мат. лит., 1986. – 232 с.
18. Карданова Е. Ю. Моделирование и параметризация тестов: основы теории и приложения / Е. Ю. Карданова. – М. : Федеральный центр тестирования, 2008. – 303 с.
19. Verhelst N.D. Modeling Sums of Binary Responses by Partial Credit Model / N. D. Verhelst. – H. H. F. M. Verstalen, Measurement and Research Department Reports 97–7. – Cito, Arnhem, 1997.
20. Muraki E. A. Generalized Partial Credit Model: Application of an EM Algorithm / E. A. Muraki // Applied Psychological Measurement. – 1992. – No 16(2). – P. 159–176.
21. Елисеев И. Н. Теоретические основы алгоритма расчета латентних переменных программным комплексом RILP – 1M / И. Н. Елисеев // Программные продукты и системы. – 2011. – № 2. – С. 67–72.

22. Cohen L. Approximate Expression for Parameter Estimates in the Rasch Model / L. Cohen // British Journal of Mathematical and Statistical Psychology. – 1979. – Vol. 32, No 1. – P. 113–120.
23. Wright B.D. Best Test and Self-Tailored Testing / B. D. Wright, G. A. Douglas. – Research Memorandum No 19. – Chicago : MESA Press, 1975 [Electronic resource]. – Access mode : www.rasch.org/rmt/rmt102q.htm/.
24. Linacre J. M. PROX with Missing Data, or Known Item or Person Measures / J. M. Linacre // Rasch Measurement Transactions. – 1994. – Vol. 8, [Electronic resource]. – Access mode : www.rasch.org/rmt/rmt122q.htm/.
25. Linacre J. M. Estimating Rasch Measures with Known Polytomous Item Difficulties / J. M. Linacre // Rasch Measurement Transactions. – 1998. – Vol. 12, [Electronic resource]. – Access mode : www.rasch.org/rmt/rmt122q.htm/.
26. Linacre J. M. Dichotomous Mean-square Chi-square fit statistics / J. M. Linacre, B. D. Wright // Rasch Measurement Transactions. – 1994. – Vol. 8, No 2. – P. 360.
27. Wright B. D. Computation of OUTFIT and INFIT Statistics / B. D. Wright, G. N. Masters // Rasch Measurement Transactions. – 1990. – Vol. 3, [Electronic resource]. – Access mode : www.rasch.org/rmt/rmt34e.htm/.
28. Linacre J. M. PROX for Polytomous Data / J. M. Linacre // Rasch Measurement Transactions. – 1995. – Vol. 8, [Electronic resource]. – Access mode : www.rasch.org/rmt/rmt84q.htm/.
29. Hambleton R. K. Fundamentals of Item Response Theory / R. K. Hambleton, H. Swaminathan, H. J. Rogers. – Sage Publications Inc. London, 1991, Chapter 2, p. 9–12 and Exercise 6. – P. 29–31.
30. Tristan A. Chi-square Local Independence Meets the Rasch Model / A. Tristan // Rasch Measurement Transactions. – 2002. – Vol.16, No 1. – P. 861.
31. Rizopoulos D. Ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses / D. Rizopoulos // Journal of Statistical Software. – Nov. 2006. – Vol. 17, Issue 5 [Electronic resource]. – Access mode : http://www.jstatsoft.org/.
32. Латентно-структурный анализ и теория тестов / Математические методы в социальных науках // Под ред. П. Лазарсфельда, Н. Генри ; пер. с англ. – М. : Прогресс, 1973. – 297 с.
33. Bock R. D. Full-Information Item Factor Analysis / R. D. Bock, R. D. Gibbons, E. Muraki // Applied Psychological Measurement. – 1988. – No 12. – P. 261–280.
34. Gibbons R. D. Full-information item bi-factor analysis / R. D. Gibbons, D. Hedeker // Psychometrika. – 1992. – Vol. 57. – P. 423–436.
35. Vermunt J. K. Latent Class Cluster Analysis / J. K. Vermunt, J. Magidson // J. A. Hagenaars, A. L. McCutcheon (eds.). Advances in Latent Class Analysis. – Cambridge University Press, 2002.
36. Glas C. A. A Bayesian Approach to Person Fit Analysis in Item Response Theory Models / C. A. Glas, R. R. Meijer // Applied Psychological Measurement. – 2003. – Vol. 27, No 3. – P. 217–233.
37. Bayesian Modeling Using WinBUGS. – A John Viley&Sons, Inc, publication, 2009. – 506 p.
38. Wim J. van der Linden. Elements of Adaptive Testing / Wim J. van der Linden and Cees A. W. Glas. –New York-Dordrecht-Heidelberg-London: Springer, 2010. – 437 p.
39. Авраменко О. В. Моделі та методи статистичної обробки результатів тестування: огляд монографій та підручників / О. В. Авраменко // Наукові записки НДУ ім. М. Гоголя. Психолого-педагогічні науки. – 2011. – № 10. – С. 17–24.
40. Bollen K.A. Latent Variables in Psychology and Social Scinces / K. A. Bollen // Annu. Rev. Psychol. – 2002. – Vol. 53. – P. 605–634.
41. Скрытая марковская модель. – Википедия [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Скрытая марковская модель.
42. Эфрон Б. Нетрационные методы многомерного статистического анализа / Сборник статей // Пер. с англ. Ю. П. Адлера, Ю. А. Кошевника, В. Н. Солнцева ; под ред. Ю. П. Адлера. – М. : Финансы и статистика, 1988. – 264 с.

Н. Е. СИНИЦКИЙ,
кандидат физ.-мат. наук, доцент,
доцент кафедры информационных систем и технологий
Национальная академия статистики, учета и аудита

Статистические инструменты измерения качества образования. Часть 3. Современный поход

Представлен обзор статистических основ тестологии. Описаны базовые задачи, математические модели и основные расчетные формулы современной (IRT) теории тестирование, позволяющее из статистических позиций оценить правильность построения, разрешение, стандартную ошибку и надежность тестовых измерений.

Ключевые слова: *классическая теория тестирования, испытуемые, оценки, тест, IRT-модель, PL-модель, PCM-модель, шкала Раша, логит-оценка.*

M. YE. SINYTSKYI,
PhD (Phys.-Math.), Associate Professor,
Associate Professor of Department for Information Systems and Technologies,
National Academy of Statistics, Accounting and Audit

Statistical Tools for Measuring the Quality of Education Part 3. Classical Approach

The article presents an overview of the statistical grounds of testology. The purpose of this paper is to explain the inexperienced readers, such as teachers of economic disciplines, opportunities of improvement of quality of education with the utilization of objective and impartial tools of students' achievement measurement, known as tests.

The first part of the article overviews the shortcomings of the traditional system of evaluation of educational achievements that is built around the use of ordinal scales. The limitations imposed to the possibilities of statistical processing of the raw data by the type of scale are shown. Basic tasks, the corresponding mathematical models, statistical characteristics and sub-test score evaluation reliability formulas are described.

The second part of the article describes approaches to the determination of reliability, uniformity and resolution of the test, built on the analysis of the correlation between students' answers to the identical questions asked. Options of conversion of primary points to a quantitative scale are provided. Ways of lowering the probability of correctly guessed answers are shown. The approach to processing of results of complex test structures is given and the possibility of utilization of two-factor analysis of variance (2 Way ANOVA) for dichotomous tests reliability estimation is demonstrated.

The third and the fourth parts of the article are devoted to the modern theory of tests (IRT).

The third part provides an analysis of shortcomings of the CTT, which were the main focus of efforts to overcome of IRT supporters during the last 60 years. The theoretical basis for building a Rach model and its subsequent developments is described. The methodology of estimation of properties of the test by its characteristic curves and parameters of its information function is illustrated. The basic equation, the correspondent solution of which gives an estimate of the probability of obtaining a certain personal score of a test is formulated.

The fourth part of the article provides various methods of finding a solution of the basic equation for the 1PL and 2PL – models and data preparation for a correct use. Several software packages, both considered to be classical tools as well as brand new ones, are overviewed. An example of ranking of NASOA students' achievements obtained by traditional evaluation and IRT approach is given.

Keywords: *quality education, testing, scaling, observed scores, correlation, reliability, resolution, uniformity, latent variable model of Rush, characteristic curves, logit.*