УДК 001.891:371.263

М. Є. СІНИЦЬКИЙ,
кандидат фізико-математичних наук, доцент,
доцент кафедри інформаційних систем і технологій,
Національна академія статистики, обліку та аудиту

Статистичні інструменти вимірювання якості освіти. Частина 4. Сучасний 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.*

Відомо декілька комп`ютерних алгоритмів знаходження параметрів *IRT*-моделей. Огляд їх сучасних реалізацій зайняв би багато місця, тому зупинимось на тих із них, що широко відомі та можуть бути реалізовані на рівні електронних таблиць (ЕТ).

Запис (15) для всіх випробовуваних у тесті дає систему нелінійних алгебраїчних рівнянь (*CHAP*), що в принципі розв`язується методом зваженого (узагальненого) методу найменших квадратів. Але цей шлях натикається на виродженість інформаційної матриці, і мова може йти лише про нормальне псевдорішення задачі найменших квадратів [15].

Як інший варіант можна побудувати *CHAP*, використовуючи метод моментів [4]. Він був запропонований ще К. Пірсоном, який довів, що прирівнювання маточікувань первинних балів (13) і (14) самим первинним балам дає достатньо хороші та незміщені оцінки. Для *PCM*-моделі (42) така *CHAP* має вигляд:

$$\begin{cases} y_i - \sum_{j=1}^{M}\sum_{k=1}^{l_j}(k \cdot P_{ijk}) = 0; \quad 0 \le y_n \le G; \quad y_i = i \\ N_{jg} - \sum_{i=0}^{G}(N_i \cdot \sum_{k=g}^{l_j} P_{ijk}) = 0 \quad 1 \le j \le M; \quad 1 \le g \le l_j \end{cases}, \tag{50}$$

де y_i – персональний бал випробовуваних, які складають категорію i;

P_{ijk} – умовна ймовірність того, що випробуваний при виконанні j-го завдання набере i балів, тобто йдеться про підстановку у (50) формули (42);

M – кількість завдань у тесті;

l_j – максимально можливий бал за j-те завдання;

N_{jg} – кількість випробовуваних, які набрали за j-те завдання не менше g балів;

N_i – кількість випробовуваних, бал яких за тест дорівнює i;

G – максимальна можлива оцінка за тест:

$$G = \sum_{j=1}^{M} l_j . \tag{51}$$

$\sum_{k=g}^{l_j} P_{ijk}$ – імовірність того, що випробуваний з категорії i набере за виконання j-го завдання не менше, ніж g балів [4].

Система (50) містить $2G+1$ нелінійне рівняння і при $\mathbf{a}_j$**=1** стільки ж змінних. Для її розв`язання застосовується метод торканих [10].

Л. Коеном для знаходження параметрів *1PL*-моделі було розроблено т. зв. *PROX*-алгоритм (*Normal Approximation Estimation Algorithm*) [20]. Із його використанням, якщо матриця первинних відповідей не має пропусків, результату досягають за одну ітерацію за формулами:

$$\theta_n^{(1)} = \bar{\theta} + a_\theta \cdot \theta_n^{(0)}; \quad \beta_j^{(1)} = \bar{\beta} + a_\beta \cdot \beta_j^{(0)}; \tag{52}$$

де $\theta_n^{(0)}$ і $\beta_j^{(0)}$ – нульові наближення, відповідно, рівня підготовленості випробовуваного та трудності j-го завдання:

$$\theta_n^{(0)} = \ln(\frac{\bar{y}_n}{1-\bar{y}_n}), \quad \beta_j^{(0)} = \ln(\frac{1-\bar{w}_j}{\bar{w}_j}) \ , \tag{53}$$

де $\bar{y}_n$ – середні арифметичні персональні бали n-го випробовуваного;
$\bar{w}_j$ – середній арифметичний індивідуальний бал j-го завдання;
$\bar{\theta}$ і $\bar{\beta}$ – середні арифметичні величин (53):

$$\bar{\beta} = \frac{1}{N} \cdot \sum_{j=1}^{M} \beta_j^{(0)}; \quad \bar{\theta} = \frac{1}{M} \cdot \sum_{n=1}^{N} \theta_n^{(0)} \ , \tag{54}$$

a_θ і a_β – коефіцієнти зростання, відповідно, рівня підготовленості випробовуваного та трудності завдання:

$$a_\theta = \sqrt{\frac{1+\frac{D_\beta}{2{,}89}}{1-\frac{D_\theta \cdot D_\beta}{8{,}35}}}; \quad a_\beta = \sqrt{\frac{1+\frac{D_\theta}{2{,}89}}{1-\frac{D_\theta \cdot D_\beta}{8{,}35}}}; \tag{55}$$

D_β і D_θ – вибіркові оцінки дисперсій, відповідно, величин β і θ:

$$D_\beta = s_\beta^2 = \frac{1}{N-1} \cdot \sum_{j=1}^{M} (\beta_j^{(0)} - \bar{\beta})^2; \quad D_\theta = s_\theta^2 = \frac{1}{N-1} \cdot \sum_{n=1}^{N} (\theta_n^{(0)} - \bar{\theta})^2. \tag{56}$$

При цьому якість результатів $\theta_n^{(1)}$ і $\beta_j^{(1)}$ визначають стандартні помилки:

$$s_{\theta_n^{(1)}} = SE(\vartheta_n^{(1)}) = \frac{a_\theta}{\sqrt{\frac{Y_n}{M} \cdot (M - Y_n)}}; \quad s_{\beta_j^{(1)}} = SE(\beta_j^{(1)}) = \frac{a_\beta}{\sqrt{\frac{w_j}{N} \cdot (N - w_j)}} \ . \tag{57), 58}$$

Статистичне моделювання показало, що точніші результати дає застосування методу максимізації логарифмічної функції правдоподібності (*Joint Maximum Likelihood Estimation*, або *JMLE*):

$$\ln L(Y \mid \theta_n, \text{інші параметри}) \rightarrow max \ . \tag{59}$$

Попередній етап будь-якого алгоритму пошуку параметрів Раш-моделей представляє процедуру відбору завдань, а саме: аналіз первинних даних на відповідність умовам можливості застосування моделей Раша. Задача – упорядкування матриці відповідей та пошук і вилучення з неї неінформаційних рядків і стовпців, а також таких, що містять "збурення" типу наявності відповідей на більш важкі завдання та "провалів" простіших завдань. Причини цього явища аналізувались у [1]. Завдання, відповіді на які слабо корелюють з персональним балом і вилучення яких приводить до зростання коефіцієнта альфа Кронбаха (див. формулу (54) у [2]), мають бути вилучені з тесту. Тобто в цьому плані *IRT* не відрізняється від *CTT*, хоча новітні підходи не завжди вимагають такої підготовки.

Калібрування завдань, тобто визначення порогових значень категорій оцінок з урахуванням умов параметризації відбувається у рамках загального ітераційного процесу. Але після отримання результатів слід проаналізувати розподіл завдань за трудністю впродовж *RM*-шкали. Порожнечі (вікна) більш широкі, ніж (1,0–1,5) логіта, є неприпустимими особливо в середині шкали, де точність вимірювання є найвищою. Тобто потрібно заповнювати ці вікна додатковими завданнями.

Рекурентні формули обчислення параметрів моделі отримують з ньютонівського розв'язку нелінійного рівняння $L(x) = 0$, що покладено в основу методу Ньютона – Рафсона (див., напр., [15; 19]):

$$x^{(v+1)} = x^{(v)} - \frac{L(x^{(v)})}{L'(x^{(v)})} \ , \tag{60}$$

де $L(x)$ – функція правдоподібності, яка має бути двічі диференційована в околиці точки $x = 0$;

x – шукані параметри моделі: θ, β тощо;

v – номер ітерації.

Наприклад, для *1PL* рекурентні формули мають вигляд [21]:

$$\theta_n^{(v+1)} = \theta_n^{(v)} - \frac{Y_n - \sum_{j=1}^{M} P_{nj}^{(v)}}{\sum_{j=1}^{M} [P_{nj}^{(v)} \cdot (1 - P_{nj}^{(v)})]}, \quad \text{або} \quad \theta_g^{(v+1)} = \theta_g^{(v)} - \frac{g - \sum_{j=1}^{M} \hat{P}_{gj}^{(v)}}{\sum_{j=1}^{M} [\hat{P}_{gj}^{(v)} \cdot (1 - \hat{P}_{gj}^{(v)})]}, \tag{61}$$

$$\beta_j^{(v+1)} = \beta_j^{(v)} - \frac{-w_j + \sum_{n=1}^{N} P_{nj}^{(v)}}{\sum_{n=1}^{N} [P_{nj}^{(v)} \cdot (1 - P_{nj}^{(v)})]} \cong \beta_j^{(v)} - \frac{-w_j + \sum_{g=0}^{M} (P_{gj}^{(v)} \cdot N_g)}{\sum_{g=0}^{M} [P_{gj}^{(v)} \cdot (1 - P_{gj}^{(v)}) \cdot N_g]}, \tag{62}$$

де Y_n – персональний бал n-го випробовуваного за формулою (8);

w_j – первинний бал j-го завдання за формулою (9);

g – персональний бал, або індекс категорії випробовуваних чисельністю N_g.

Для моделі рейтингової шкали маємо [22–25]:

$$\theta_n^{(v+1)} = \theta_n^{(v)} - \frac{y_n - \sum_{j=1}^{M} \sum_{g=0}^{l_j} (g \cdot P_{njg}^{(v)})}{\sum_{j=1}^{M} \left\{ \sum_{g=0}^{l_j} (g^2 \cdot P_{njg}^{(v)}) - [\sum_{g=0}^{l_j} (g \cdot P_{njg}^{(v)})]^2 \right\}}, \tag{63}$$

$$\beta_j^{(v+1)} = \beta_j^{(v)} - \frac{-w_j + \sum_{n=1}^{N} \sum_{g=0}^{l_j} (g \cdot P_{njg}^{(v)})}{\sum_{n=1}^{N} \left\{ \sum_{g=0}^{l_j} (g^2 \cdot P_{njg}^{(v)}) - [\sum_{g=0}^{l_j} (g \cdot P_{njg}^{(v)})]^2 \right\}}, \tag{64}$$

де позначення відповідають позначенням у формулі (42).

$$d_g^{(v+1)} = d_g^{(v)} + \ln\left(\frac{N_{g-1}^{(v)}}{N_g^{(v)}}\right) + \ln\left(\frac{\hat{N}_{g-1}^{(v)}}{\hat{N}_g^{(v)}}\right), \tag{65}$$

де $N_g^{(v)}$ і $\hat{N}_g^{(v)}$ – відповідно, спостережувана та теоретично очікувана кількості випробовуваних, персональний бал яких склав g.

В якості нульового наближення для запуску ітераційної процедури крім нульових значень зазвичай обирають величини, розраховані за формулою (54), а при використанні частотного підходу – формули (19):

$$\theta_g^{(0)} = \ln\left(\frac{\bar{y}_g}{1 - \bar{y}_g}\right), \tag{66}$$

де $\bar{y}_g$ – середній арифметичний персональний бал у категорії випробовуваних чисельністю N_g, що набрали g балів.

Крім того, на кожному кроці з отриманого значення $\beta_j^{(v+1)}$ віднімається, $\bar{\beta}^{(v+1)}$ що забезпечує умову параметризації (17).

Для покращення збігу процедури автори [21] запропонували нульове наближення $\theta_n^{(0)}$ для *1PL*-моделі обирати як *PROX*-оцінку Коена, що дорівнює [20]:

$$\theta_n^{(0)} \approx \overline{\beta}^{(0)} + \sqrt{(1 + s^2(\beta^{(0)})/2{,}89} \cdot \ln\left(\frac{Y_n}{M - Y_n}\right), \quad (67)$$

де Y_n – персональний бал n-го випробовуваного, отриманий за виконання M завдань; при цьому:

- якщо реально $Y_n = 0$, то у розрахунках приймають $Y_n = 0{,}5$;
- якщо реально $Y_n = M$, то у розрахунках приймають $Y_n = M\text{-}0{,}5$;

$\overline{\beta}^{(0)}$ – нульове наближення середньої арифметичної трудностей тестових завдань, обчислене за формулою (57) у логітах;

$s^2(\beta^{(0)})$ – вибіркова оцінка дисперсії нульового наближення трудності завдань тесту:

$$s^2(\beta^{(0)}) = \frac{1}{M-1} \cdot \sum_{j=1}^{M} (\beta_j^{(0)} - \overline{\beta}^{(0)}), \quad (68)$$

Автор [22] спростив процедуру, запропонувавши для $\theta_n^{(0)}$ оцінку:

$$\theta_n^{(0)} = \overline{\beta} + \ln\left(\frac{Y_n - Y_n^{(\min)}}{Y_n^{(\max)} - Y_n}\right), \quad (69)$$

де $Y_n^{(\min)}$ і $Y_n^{(\max)}$ – відповідно, мінімальний та максимальний персональний бал серед всіх випробовуваних;

$\overline{\beta}$ – середня складність завдань згідно з (54).

Щоб подолати можливий незбіг ітераційного процесу, на кожному його кроці починаючи з другого пропонується контролювати виконання нерівності:

$$abs(\theta_n^{(v+1)} - \theta_n^{(v)}) > abs(\theta_n^{(v+2)} - \theta_n^{(v)}), \quad (70)$$

Якщо зміни в очікуваних значеннях персональних балів не зменшуються, а, навпаки, збільшуються, знаменник (*denom*) у формулі (61) обирають як:

$$denom = \max[denom \times 2;\ 1{,}0], \quad (71)$$

Крім того, не рекомендовано змінювати оцінку більше, ніж на один логіт від її значення на попередній ітерації:

$$\theta_n^{(v+1)} = \max[\min(\theta_n^{(v)} + 1; \theta_n^{(v+1)}); \theta_n^{(v)} - 1], \quad (72)$$

Ітераційний процес закінчується, коли різниця $\theta_n^{(v+1)} - \theta_n^{(v)} = \Delta\theta_n^{(v+1)}$ досягне малої величини, наприклад, 0,01 логіта.

Останнім етапом є перевірка якості отриманої моделі. В *IRT*-теорії застосовано підхід до вимірювання, що отримав назву *model based measurement* – не модель повинна відповідати емпіричним даним, а дані повинні відповідати моделі. Тобто, як вже згадувалося, для оцінювання патентного фактора слід використовувати тільки ті завдання, що відповідають цій моделі вимірювання. Решта завдань повинні бути змінені або відбраковані [7].

Аби перевірити, чи відповідають певні тестування тій або іншій *IRT*-моделі, використовують т. зв. *fit*-статистики (статистики відповідності, див. напр., [26; 27]). Найчастіше вони мають вигляд:

$$V(\theta_n) = \sum_{j=1}^{M} e_{nj}^2 v_j(\theta_n) = \sum_{j=1}^{M} [y_{nj} - P_{nj}(\theta)]^2 v_j(\theta_n), \quad (73)$$

де e_{nj} – різниця (залишок) між спостережуваною оцінкою за відповідь n-го випробовуваного на j-те завдання тесту та її очікуваним (модельним) значенням;

$v_j(\theta)$ – статистична вага j-го залишку.

Кінцеві цілі *fit*-статистик – визначити початкові дані, що не відповідають моделі. В роботі [27] наведено приклад використання статистики (69) для аналізу відповідності даних *IRT*-моделі.

Залишки (точкові абсолютні помилки) оцінювання зазвичай представляють як:

$$e_{nj} = y_{nj} - E(y_{nj}), \quad (74)$$

де y_{nj} – оцінка за відповідь n-го випробовуваного на j-е завдання;
$E(y_{nj})$ – математичне очікування оцінки y_{nj} згідно з обраною *IRT*-моделлю.
Якщо в якості статистичної ваги у формулі (69) використати величину:

$$v_j(\theta_n) = 1/\sqrt{Var(e_{nj})} \quad , \tag{75}$$[1]

то статистика (69) буде наближатися χ^2-розподілом Пірсона. В *IRT*-літературі вона має назву *OUTFIT*-статистики, а її складові – стандартизованих залишків:

$$z_{nj} = \frac{e_{nj}}{\sqrt{D(e_{nj})}} = \frac{e_{nj}}{\sqrt{(1/N)\cdot\sum_{n=1}^{N} e_{nj}^2}} \quad , \tag{76}$$

Величина z_{nj} має стандартний нормальний розподіл з маточікуванням $E(z_{nj}) = 0$ і

$$E(z_{nj}^2) = 1 \text{ і } Var(z_{nj}^2) = [Kurt(e_{nj}) / D^2(y_{nj})] - 1 \quad , \tag{77}$$

$Var(z_{nj}) = 1$. Відповідні параметри розподілу величини дорівнюють:

де $Var(y_{nj}) = \sum_{y_{nj}=0}^{m_j} \{[y_{nj} - E(y_{nj})]^2 v_j(\theta_n)\}$ – варіація величини y_{nj}; (78)

$$E(y_{nj}) = \sum_{y_{nj}=0}^{m_j} y_{nj} v_j(\theta_n) \quad , \tag{79}$$

$Kurt(e_{nj})$ – ексцес розподілу величини e_{nj};
m_j – кількість питань в j-му завданні.
Середній квадрат величини e_{nj}, тобто середнє арифметичне залишків відповідей всіх випробовуваних на j-те питання (*OUTFIT MEANSQ*)? розраховується за формулою:

$$OM_j = \frac{1}{N}\cdot\sum_{n=1}^{N}\frac{e_{nj}^2}{Var(e_{nj})} = \sum_{n=1}^{N}\frac{z_{nj}^2}{N} \quad , \tag{80}$$

При цьому за нормального розподілу *OM*:

$$E(OM_j) = 1;\ Var(OM_j) = MSWD^2(OM_j) = \sum_{n=1}^{N}[Kurt(e_{nj})/D^2(y_{nj})]/N^2 - 1/N, \tag{81}$$[2]

За допомогою т. зв. перетворення Вілсона – Хілферті отримують стандартизований середній квадрат величини e_{nj} (*OUTFIT ZSTD*):

$$OZ_j = (\sqrt[3]{OM_j} - 1)\cdot\left(\frac{3}{MSWD(OM_j)}\right) - \left(\frac{MSWD(OM_j)}{3}\right) \quad , \tag{82}$$

При цьому $E(OZ_j) = 0$; $Var(OZ_j) = 1$.
Щоб зменшити вплив малоінформативних з низькою дисперсією повзцільових відповідей, критерії (80) і (82) додатково зважують, отримуючи т. зв. *INFIT*-статистики:
INFIT MEANSQ:

$$IM_j = \sum_{n=1}^{N}\varepsilon_{nj}^2 / \sum_{n=1}^{N} Var(\varepsilon_{nj}) \quad , \tag{83}$$

При цьому за нормального розподілу *IM*:
$E(IM_j) = 1$;

$$Var(IM_j) = MSWD^2(IM_j) = \sum_{n=1}^{N}[Kurt(e_{nj}) - Var^2(y_{nj})]/N^2 - 1/N, \tag{84}$$

[1] Var – скорочення від variation.
[2] MSWD (Mean square weighted deviation) – зважене середньоквадратичне відхилення оцінки за j-те завдання.

INFIT ZSTD:

$$IZ_j = (\sqrt[3]{IM_j} - 1) \cdot \left(\frac{3}{MSWD_j} \right) - \left(\frac{MSWD_j}{3} \right), \tag{85}$$

Імовірнісну інтерпретацію наведених статистик згідно з [28] наведено в табл. 1 і 2.

Таблиця 1

Інтерпретація OUTFIT- й INFIT-статистик у варіанті MEANS

Значення	Інтерпретація
> 2	Завдання порушують систему вимірювань. Допустимими є 1–2 таких завдання
1,5–2,0	Завдання малопродуктивне для вимірювання, але може бути використано без редагування
0,5–1,5	Завдання може бути використане для вимірювання
< 0,5	Малопродуктивне завдання. Може помилково породити відчуття високої надійності завдань

Таблиця 2

Інтерпретація *OUTFIT*- й *INFIT*-статистик у варіанті *ZSTD*

Значення	Інтерпретація
> 3	Дані не вписуються в модель або дуже малий обсяг вибірки
2,0–2,9	Дані мало передбачувані
-2 – +2	Дані добре передбачувані
< -2	Дані дуже передбачувані

Для моделі *1PL*:

$$E(y_{nj}) = a_{nj}, \tag{86}$$

або за формулою (19)

$$E(y) = \hat{P}_{yj} = N_{yj} / N_y \tag{87}$$

і

$$z_{nj} = \frac{a_{nj} - P_{nj}}{\sqrt{P_{nj}(1 - P_{nj})}} = (2a_{nj} - 1) \cdot \exp[(2a_{nj} - 1) \cdot \frac{\beta_j - \theta_n}{2}] \tag{88}$$

або у відповідності до формули (20):

$$z_{yj} = \frac{y - \hat{P}_{yj}}{\sqrt{\hat{P}_{yj} \cdot (1 - \hat{P}_{yj}) / (N_y - 1)}} = \frac{y \cdot N_y - N_{yj}}{\sqrt{N_{yj} \cdot (N_y - N_{yj}) / (N_y - 1)}} \tag{89}$$

де у формулах (86) і (88) P_{nj} – оцінка ймовірності отримання n-м випробовуваним за j-те питання 1 балу;

а у формулах (87) і (89) – оцінка ймовірності отримання всіма випробовуваними за j-те питання y балів.

N_j – кількість випробовуваних, які отримали оцінку j.

Fit-статистика (69) для цієї моделі має вигляд:

$$\sum_{j=1}^{M} \exp[(2a_{nj} - 1) \cdot (\beta_j - \theta_n)] = \sum_{j=1}^{M} \chi_j^2 = \chi_{M-1}^2, \tag{90}$$

або

$$\sum_{y=0}^{M} \frac{N_y \cdot (\hat{P}_{yj} - P_{yj})^2}{P_{yj}} = \sum_{j=1}^{M} \frac{(N_{yj} - N_y \cdot P_{yj})^2}{N_{yj}} = \sum_{j=1}^{M} \chi_j^2 = \chi_{M(M+1)}^2, \tag{91}$$

$$\sum_{n=1}^{N} \exp[(2a_{nj} - 1) \cdot (\beta_j - \theta_n)] = \sum_{n=1}^{N} \chi_j^2 = \chi_{N-1}^2 \tag{92}$$

де позначення відповідають позначенням у формулі (19), а нижній індекс у позначенні критерію χ^2 означає число степенів свободи.

До речі, число степенів свободи χ^2-розподілу у формулі (91) дорівнює $v = h$-1, де h – кількість груп, на які розбиваються випробовувані в залежності від набраного персонального балу. Тому оскільки $h=M+1$, то $v=M$. Для – статистики у формулі (91) $v=M(M+1)$.

Нульова статистична гіпотеза H_0: генеральна сукупність випробовуваних і тестових завдань є такою, що ймовірність P_{xj} адекватно моделюється формулою Раша (22), підтверджується, якщо $\chi^2_{спост} \le \chi^2_{крит}(\nu)$, де $\chi^2_{крит}(\nu)$ визначають за таблицями Пірсона, розв'язуючи рівняння $P(\chi^2_{спост} > \chi^2_{крит}) = 1-\alpha$, де α – рівень значимості (зазвичай α=0,05).

Автори [27] конкретизували фактори *fit*-статистик для політомичної *Rating Scale* моделі як:

$$E(y_{nj}) = \sum_{g_j=0}^{m_j} g_j \cdot v_{njg} \ , \tag{93}$$

де g_j – кількість балів, набраних за виконання j-го завдання; $g_j = 0, 1, \ldots, m_j$;

$$v_{njg} = \exp[\sum_{h=0}^{g_j}(\theta_n - \delta_{jh})] / \sum_{g_j=0}^{m_j}\exp[\sum_{h=0}^{g}(\theta_n - \delta_{jh})] \ ; \tag{94}$$

$$Var(y_{nj}) = \sum_{g_j=0}^{m_j}[g_j - E(y_{nj})]^2 \cdot v_{njg} \ ; \tag{95}$$

$$Kurt(y_{nj}) = \sum_{g_j=0}^{m_j}[g_j - E(y_{nj})]^4 \cdot v_{njg} \ . \tag{96}$$

Підстановкою виразу (94) у формулу (95) можна отримати розрахункову формулу для – статистики адекватності j-го завдання *Rating Scale* моделі.

Fit Analysis часто доповнюють порівнянням двох гістограм розподілу параметрів θ і β (*Item-person Map*). Зона їх збігу на *RM*-шкалі визначає порогову величину θ для нормативно-орієнтованих тестів. Додатне значення різниці середніх значень вказує на ступінь легкості, а від'ємне – трудності тесту для випробовуваних.

Наведені *fit*-статистики зрештою пояснюють, наскільки завдання тесту узгоджені в плані вимірювання однієї безперервної латентної змінної θ або вимірюються різні фактори.

Піонери *IRT* також підтримували ідею CTT, що локальна незалежність завдань може бути перевірена за процедурою розрахунку критерію χ^2 [26; 29], що полягає в побудові таблиці суміжності 2x2 (табл. 3), яка складена з частот відповідей на всі можливі пари завдань, надані випробовуваними, які набрали однакові персональні бали[3].

Таблиця 3

Таблиця суміжності частот відповідей на *i*-те та *j*-те завдання у випробовуваних, які набрали однаковий персональний бал *Y*

	Відповіді на завдання *j*	
Відповіді на завдання *i*	Правильні	Неправильні
Правильні	*A*1	*B*1
Неправильні	*A*2	*B*2

Розрахункова формула для критерію $\chi^2(1)$ має вигляд:

$$\chi^2 = \frac{(A1+B1+A2+B2)\cdot(A1\cdot B2 - B1\cdot A2)}{(A1+B1)\cdot(A2+B2)\cdot(A1+A2)\cdot(B1+B2)} \ . \tag{97}$$

де позначення відповідають табл. 3.

[3] Процедура представлена в багатьох статистичних пакетах, наприклад, Statistica 6.0.

Недоліки цього підходу було продемонстровано у роботі [30]. Натомість за аналогією з табл. 3 почали порівнювати за критерієм χ^2 дво- і тристоронні стандартизовані залишки виду (76), проводячи групування за однаковими оцінками за відповіді на пари чи, відповідно, трійки завдань [31].

Окрім *fit*-статистик розмірність *RM*-шкали визначають шляхом порівняння відношення першого до другого власних значень для кожної масштабованої матриці чотирьохкоміркових кореляцій. Свідченням одновимірності є те, що на перший фактор припадає значна частка (≥80%) дисперсії [8]. У деяких програмах до цієї операції залучають *SVD*-розкладання[4] матриці залишків.

Для побудови *IRT*-моделей було розроблено багато спеціалізованих комп`ютерних пакетів. Серед найвідоміших:

1. Програмні продукти *WINSTEPS* (ознайомча версія – *MINISTEPS*), *BIGSTEPS*, *BICAL*, створені J. M. Linacre і L. Cohen у рамках досліджень, що виконувались у Чиказькому університеті під керівництвом B. D. Wright, M. H. Stone, G. Masters, (www.winsteps.com).

2. Ліцензійна діалогова система *RUMM* (*Rasch Unidimensional Measurement Models*) версії: 2010, 2020, 2030 [15], розроблена під керівництвом B. A. Douglas і D. Andrich в Мердокському університеті (Австралія).

3. *Bilog –MG3*, *Multilog*, *Parscale 4* (програми оцінювання параметрів моделі Бірнбаума та її розширення для політомічних завдань) (www.assess.com; www.ssicentral.com).

4. *Facets 3.71.4* (остання версія програми для *Many-Facet Rasch Model*) (www.winsteps.com/facets.htm).

5. *Rilp-1* (Рашевський вимірник латентних змінних) – пакет від центру тестування Південно-Російського державного університету економіки та сервісу [21].

6. *IRTPRO 3 for Windows* (http://www.ssicentral.com/irt/).

IRT представляє науковий напрямок, що динамічно розвивається. Його застосовують не тільки як інструмент для побудови системи управління якістю освіти, а й як основу для соціально-економічних досліджень суспільства [32].

На сьогодні створено теорію багатовимірних *IRT*-моделей та методів їх рішень [10; 33; 34]. Але слід відмітити, що сучасні підходи до визначення латентних змінних розвиваються переважно на основі ідей латентно-структурного аналізу (*Latent Class Analisys, LCA*) з використанням алгоритмів змішаної множинної регресії та прихованих марківських процесів [35–37]. На цю тему публікується багато наукових статей та навіть навчальних посібників, резюме деяких можна знайти, наприклад, у [38–41].

Водночас поповнюється програмне забезпечення, що реалізує нові ідеї ідентифікації скритних параметрів:

1. *QUEST*, *RASCAL*, *CONQUEST*, *SAS PROCMMIXED* (включено до пакету SAS6.0 і вище), *MIXOR*, *Latent GOLD* (багатовимірні та змішані регресійні моделі) (див. наприклад, http://www.scienceplus.nl/scienceplus/main/show_ pakketten_ categorie.jsp?id=38).

2. *LTM* – пакет для середовища *R* (*R Development Core Team*) (http:// CRAN. R-Project.org/).

3. *WINMIRA*, *WinBUGS* [37] – програмне забезпечення для реалізації байєсівського оцінювання параметрів моделей з латентними параметрами.

Як приклад, відповідно до *1PL*-моделі за допомогою ET *Google Docs* ("вручну") і частково пакету *Ministeps* (ознайомча версія *Winsteps*, кількість завдань обмежена 25) було оброблено реальні дані рубіжного контролю студентів одного з київських ВНЗ (120 тестових завдань, обсяг вибірки – 24 випробовуваних). Застосування ET *Google Docs* пов`язано з можливістю забезпечення доступу користувачів до достатньо великого масиву даних, а програми *Ministeps* – з її аналітичними можливостями (більше 50 варіантів представлення результатів). Відповідні файли ET *Google Docs* можна знайти в Інтернеті за адресами: htpps://docs.google.com/ spreadsheets/d/18nR3djSEjo5SP-RX7tAq4p4K7kHb- DbXPQh2biDiAcxo/edit#gid= 1130558873 (вихідні дані) та gid= 1468543099 (результати).

[4] Сингулярне розкладання (Singular Values Decomposition, SVD) дозволяє знайти власні значення (квадрати сингулярних чисел) і власні вектори дійсної матриці даних та її коваріаційної матриці (чим замінює метод головних компонент).

Попередньо, згідно з принципами *CTT*, з розгляду було вилучено 26 неінформативних завдань (на які 16 осіб дали всі правильні відповіді та 10 осіб не дали жодної правильної відповіді). Далі після розрахунку кореляційної матриці, незважаючи на достатньо високий рівень узгодженості завдань за узагальненою формулою Спірмена-Брауна (формула (48) [2]) рівний 0,91, з подальшого оброблення було вилучено 23 завдання, що мали переважно від`ємні коефіцієнти парної кореляції з іншими завданнями, а з тих, що залишилися, було вилучено ще 13 завдань, коефіцієнт бісеріальної кореляції яких був менше 0,2. Після цього було визначено показники надійності-узгодженості завдань за CTT: χ^2 для тесту в цілому (формули (53) – (54) [2][5]), *KR-20* (формула (49) [2]), альфа-Кронбаха (засобом *SPSS*: *Анализ → Шкалирование → Анализ пригодности*) і , розрахованого за формулами (75) – (78) [2]. У результаті вилучення 47 завдань з 120 критерій χ^2 змінився з 75 (при $\chi^2_{кр}$=92) до 81 (при $\chi^2_{кр}$=75); інші коефіцієнти збільшились: *KR-20* – з 0,85 до 0,92, альфа-Кронбаха – з 0,84 до 0,91 і – з 0,49 до 0,73. Як бачимо, найчутливішим виявився коефіцієнт , що визначається за процедурою *2 Way Anova* без повторень.

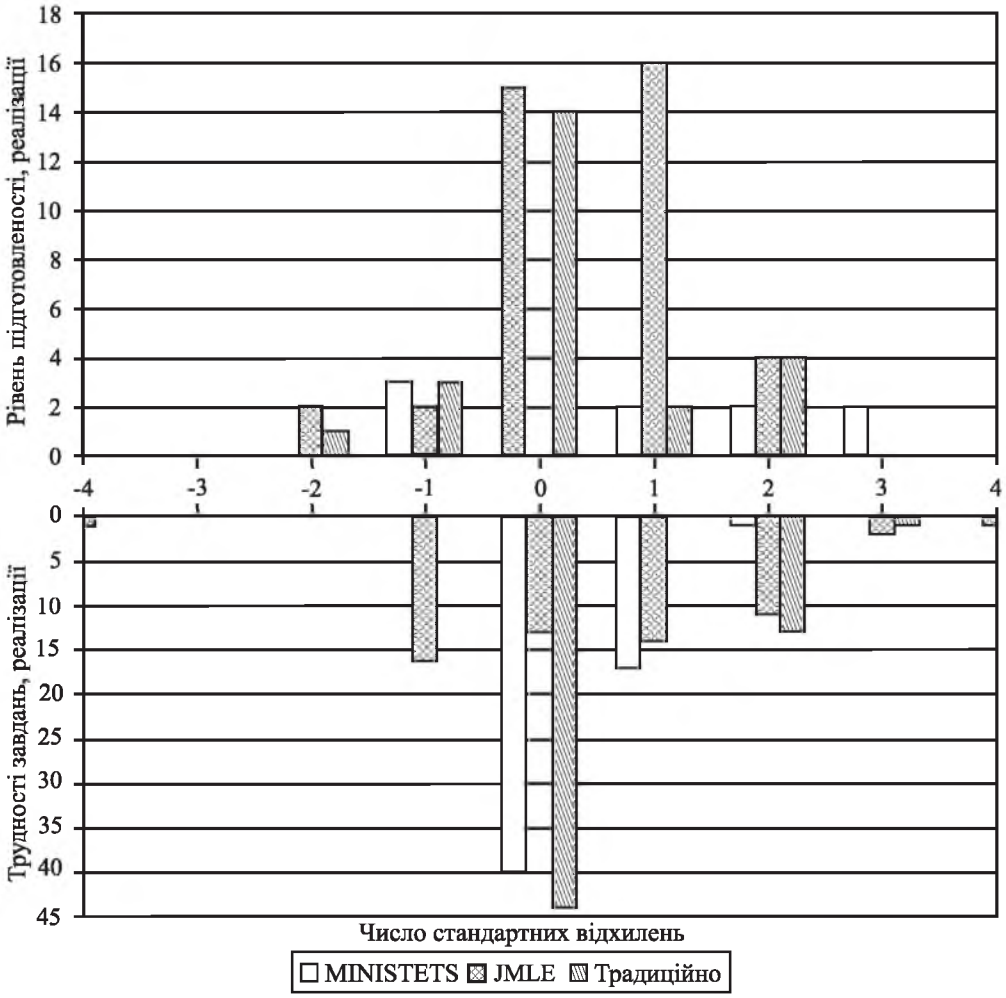

Рис. 6. Карта відповідності рівня підготовленості випробуваних і трудності завдань, оцінених у традиційний спосіб, "вручну" (алгоритм *JMLE*) і з використанням пакета *Ministeps*

Джерело: побудовано автором

[5] Як субтестп було використано непарні та парні завдання, хоча це досить умовно.

На рис. 6 наведено гістограми результатів тестування, отриманих у три способи: традиційно, як частка правильних відповідей; "вручну" за *JMLE*-алгоритмом [див. формули (59) – (62), (66) – (72)] та як усереднене значення з трьох вибірок з кроком у два завдання за допомогою програми *Ministeps 3.81.0* (оскільки більше 25 завдань за її використання обробити неможливо). Для забезпечення можливості порівняння вихідні бали трансформувались спочатку у Z-шкалу, потім у 100-бальну ss середнім 50 і дисперсією 10 балів, і нарешті у шкалу *ICTS*. Розподіл реалізацій представлено у шкалі стандартних відхилень.

З рис. 6 можна бачити, що гістограми, отримані в традиційний спосіб і з використанням пакета *Ministeps*, подібні одна до другої, причому розподіл трудностей завдань є далеким від нормального і вказує на нестачу нескладних завдань. Розрахунки за алгоритмом *JMLE* дають дещо відмінну картину. Максимум розподілу рівня підготовленості зміщений у бік вищих оцінок на 1 КО, що, як можна визначити з табл. 4, складає 1,12 логіта, або від 55 до 65 балів.

Розподіл трудностей завдань є значно більш пологим, хоча він вказує на певний брак нескладних завдань. Така розбіжність з результатами від *Ministeps* може бути пов'язана з тим, що у *Winsteps* враховується багато додаткових факторів, наприклад, уводяться поправки на розмір вибірки, на розподіл та дисперсію даних, видаляються аномальні спостереження тощо.

Таблиця 4

Описові статистики результатів тестування до їх стандартизації

Метод оброблення	Рівень підготовленості		Трудність завдань	
	Середнє арифм.	СКО	Середнє арифм.	СКО
Традиційний	62,78 балів	7,00 балів	27,51 балів	32,55 балів
JMLE	1,65 логіт	1,12 логіт	0,80 логіт	0,77 логіт
Ministeps	0,03 логіт	1,48 логіт	-0,47 логіт	2,50 логіт

Джерело: розраховано та побудовано автором

Для декого результати, показані *JMLE*-алгоритмом, можуть здатися більш привабливими, оскільки вони хоча і нижчі у 16 випадках у порівнянні з традиційними оцінками, але у 20 випадках перевищують результати від *Ministeps*. Але *Ministeps* має набагато більш широкі можливості, зокрема, результати наданих програмою звітів з аналізу стандартизованих залишків ("*Table 23 of Standardized Residual variance*") і *SVD*-розкладання матриці залишків (*SVD File*) показують наявність до 3-х значимих латентних факторів, що майже однаково впливають на результати тесту. Тобто шкала оцінювання зміщена, і тестові завдання (як мінімум 25, якщо брати за основу звіт з *fit*-статистики) потребують доопрацювання.

Висновки

1. *IRT* дозволяє вирішити три ключові задачі педагогічного вимірювання:
   - знайти параметри (трудність, роздільна здатність, узгодженість з метою тестування) тестових завдань;
   - знайти параметри випробовуваних (підготовленість, компетентність тощо);
   - підібрати функцію (модель), що описує зв'язок імовірності правильної відповіді випробовуваного зі значенням досліджуваної латентної змінної та параметрів завдання.
2. Щоб досягти незалежності оцінок трудності завдань від вибірки випробовуваних, що надає можливість створення банків завдань і об'єктивного порівняння випробовуваних, потрібно забезпечити відповідність завдань вимогам *IRT*-моделі.
3. Побудова банку завдань вимагатиме оброблення значних вибірок (порядку сотень випробовуваних і питань), тому має сенс використання чисельного моделювання процесу тестування (бутстреп процедур) [42].

4. *IRT*-технології дозволяють включення всіх оцінок з дисциплін учбового плану в оцінювання досягнення заданої компетенції, що може служити основою для поліпшення освітніх програм ВНЗ.
5. Як об'єкти вимірювання в *IRT* можуть використовуватися не тільки учбові досягнення студентів, а й різні аспекти виховної роботи, процеси та підпроцеси системи менеджменту якості ВНЗ.
6. *IRT* як піднапрям латентно-класового аналізу може замінити різноманітні методи, що використовують для сегментації споживачів; має сенс викладання *IRT* у курсах маркетингових дисциплін і технологій прийняття рішень.

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. – № 1 – С. 75–86.
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.appliedresearch.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 физико-математических наук, доцент, доцент кафедры информационных систем и технологий
Национальная академия статистики, учета и аудита

Статистические инструменты измерения качества образования. Часть 4. 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. E. 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 4. Modern 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.*

Посилання на статтю:

Сіницький М. Є. Статистичні інструменти вимірювання якості освіти.Частина 4. Сучасний підхід / М. Є. Сіницький // Науковий вісник Національної академії статистики, обліку та аудиту: зб. наук. праць. – 2016. – №1–2. – С. 99–112.