Міністерство освіти і науки, молоді та спорту України
Ніжинський державний університет імені Миколи Гоголя

Лісова Т.В.

Моделі та методи сучасної теорії тестів

Навчально-методичний посібник

This project has been funded with support from the European Commission. This publication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

УДК 371
ББК 74.04(4Укр)я73

Роботу виконано в рамках міжнародного проекту **«Освітні вимірювання, адаптовані до стандартів ЄС»** за програмою Європейського Союзу Темпус

Автор: *Т.В. Лісова*

Рецензент: доктор фізико-математичних наук, професор *О.В. Авраменко*

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

У посібнику розглядаються основні математичні моделі сучасної теорії тестів для завдань дихотомічного та політомічного типів, методи побудови оцінок латентних параметрів та методи дослідження відповідності емпіричних даних обраній моделі. Наводиться огляд програмного забезпечення для обробки результатів тестування у рамках основних моделей.

Посібник буде корисним студентам, магістрантам, що спеціалізуються на *Освітніх вимірюваннях*, а також викладачам вузів та працівникам установ сфери освіти, які цікавляться теорією і практикою об'єктивних вимірювань в освіті.

ISBN
ББК 74.04(4Укр)я73

ЗМІСТ

ВСТУП

Роботи данського математика Георга Раша наприкінці 50-х років XX століття дали поштовх до інтенсивного розвитку теоретичної бази тестування, результатом якого стала сучасна теорія під назвою Item Response Theory (IRT). Дослівний переклад українською мовою як «Теорія відповідей на питання» звучить примітивно і не відображає суті даної теорії, тому її частіше в україномовній літературі називають сучасна теорія тестів IRT на відміну від класичної теорії тестів. Можна також зустріти назву «Теорія параметризації педагогічних тестів».

У даному посібнику розглянуто деякі моделі та методи сучасної теорії тестування для тестів з дихотомічними та політомічними завданнями, які дозволяють продемонструвати вагомі переваги IRT у порівнянні з класичною теорією:

- Стійкість і об'єктивність оцінок параметра, що характеризує рівень підготовки опитаних. Джерелом стійкості є відносна інваріантність оцінок рівня підготовки від складності завдань.
- Стійкість і об'єктивність оцінок параметра складності завдань, їх незалежність від властивостей вибірки опитаних.
- Можливість вимірювання значень параметрів опитаних і завдань тесту за однією і тією ж шкалою, що має властивості інтервальної.

Застосування сучасної теорії тестів передбачає роботу з великими масивами даних, що важко уявити без використання спеціального програмного забезпечення. Тому у посібнику також коротко описано можливості кількох програм, призначених для обробки результатів тестування у рамках деяких моделей IRT.

Методи, що використовуються для побудови оцінок параметрів опитаних та завдань, перевірки відповідності емпіричних даних обраній математичній моделі, передбачають знайомство читача з основами математичної статистики.

1. МАТЕМАТИЧНІ МОДЕЛІ СУЧАСНОЇ ТЕОРІЇ ТЕСТІВ

1.1. Основні поняття та припущення сучасної теорії тестів

У сучасній теорії тестування вводиться основне припущення про існування деякого взаємозв'язку між спостережуваними результатами тестування і латентними (прихованими від безпосереднього спостереження) якостями випробовуваних, які виконують тест. Зазвичай ці латентні якості трактуються як здатності (здібності) випробовуваних або як їх рівні підготовки і умовно позначаються θ.

Передбачається, що кожному випробовуваному відповідає лише *одне* значення латентного параметра θ, який визначає спостережувані результати виконання тесту. Параметр може змінюватись у межах $(-\infty, \infty)$ і чим більше значення цього параметра, тим вища ймовірність правильної відповіді на питання. Ймовірність конкретної відповіді на питання тесту є монотонною та нелінійною функцією здібностей. У більшості моделей така функціональна залежність має вигляд *S*–подібної кривої (рис.1) і називається характеристичною кривою завдання.

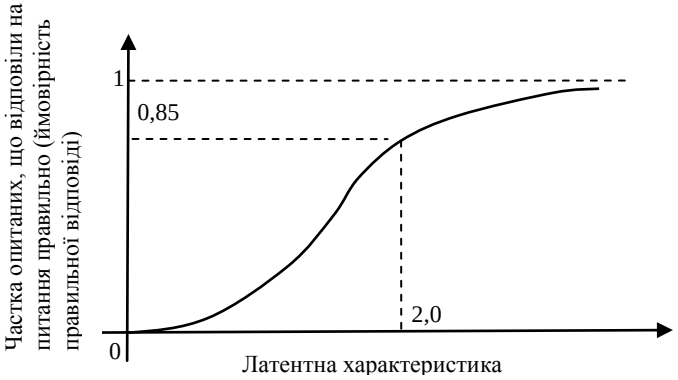


Рис.1.

Кожній точці на осі латентної характеристики θ відповідає деяка однорідна (гомогенна) підгрупа (субпопуляція) учасників, які мають однакове дане значення характеристики. Для кожної точки графіка, наприклад (2.0; 0.85), можлива така інтерпретація:

серед учасників тестування, які мають значення латентної характеристики $\theta = 2.0$, частка тих, що дали правильну відповідь на питання, становить 0.85 (або еквівалентна інтерпретація: навмання вибраний учасник із групи, яка має однакове значення латентної характеристики $\theta = 2.0$, правильно відповість на питання із ймовірністю 0.85).

Розміщення такої кривої вздовж горизонтальної осі латентної характеристики визначається складністю завдання. Чим складніше завдання, тим більше крива зміщена вправо, у сторону більших значень латентної характеристики. Для складнішого питання, щоб відповісти правильно з ймовірністю 0.85, потрібен більший рівень характеристики здатності, наприклад $\theta = 4.0$. Пологіша крива відповідає ситуації, коли учасники з великою різницею у рівнях підготовки матимуть майже однакову ймовірність правильної відповіді. Кажуть, що таке завдання має погану диференціюючу здатність. Отже, крутизна кривої залежить від здатності завдання розрізняти опитаних зі схожими рівнями підготовки. Таким чином, характеристична крива завдання є основним будівельним блоком IRT, всі інші конструкції від неї залежать.

Для демонстрації деяких теоретичних положень IRT інколи використовують характеристичну криву у вигляді ступінчастої функції, яка має стрибок при деякому значенні латентної характеристики θ^*, так, що учасники із меншими значеннями характеристики не можуть правильно відповісти на питання, а із рівними або більшими значеннями за θ^* обов'язково дадуть правильну відповідь. Але при конструюванні тестів вона рідко використовується, оскільки фактичні дані тестування найбільше сумісні саме з *S*–подібною кривою. Такою кривою, наприклад, є інтегральна функція нормального розподілу (огіва), яка широко використовувалась у ранніх дослідженнях по теорії латентних характеристик. У більшості сучасних моделей IRT використовується логістична крива, яка в усіх точках області визначення близька до нормальної огіви, але дозволяє швидше та простіше проводити різні обчислення.

Важливими припущеннями для більшості моделей сучасної теорії тестів, які будуть тут розглядатися, є *одномірність* тесту та *локальна незалежність* завдань. Тест вважається одномірним, як-

що статистична залежність між завданнями тесту може бути пояснена єдиною латентною характеристикою. Завдання тесту по всій популяції опитаних можуть бути статистично залежними, але повинна існувати єдина латентна характеристика, така, що завдання незалежні у кожній субпопуляції опитаних, яка однорідна відносно даної латентної характеристики. Іншими словами, при фіксованому значенні латентної характеристики нема ніякого зв'язку між ймовірностями правильних відповідей на різні завдання тесту. Оскільки така незалежність визначається для субпопуляції опитаних, локалізованих у єдиній точці на шкалі латентної характеристики, то її називають локальною незалежністю.

Вимога локальної незалежності завдань є суттєвою при використанні математичного апарату IRT, коли ймовірність виконання сукупності завдань знаходиться як добуток ймовірностей виконання окремих завдань. На практиці така вимога має швидше формальний характер, оскільки відповіді на питання пов'язані між собою тим більше, чим вища кореляція між питаннями тесту. Щоб забезпечити хоча б наближене відображення ідеї локальної незалежності, розробники включають у тест завдання з невисоким значенням коефіцієнта внутрішньої кореляції, відмовляються від ланцюгових завдань, коли відповідь на одне питання входить у набір даних до іншого. Зрештою, існують методи перевірки гіпотези про локальну незалежність, наприклад, тест Йєн Q_3.

Вимога одномірності не носить, як правило, суперечливого характеру, оскільки логіка розробника тесту часто слідує саме такому зразку. Він висуває гіпотезу про те, наприклад, що створюваний тест покликаний виміряти рівень підготовки з предмету, а не швидкість читання або розуміння. Однак ця вимога істотно знижує можливості одномірної IRT в тій ситуації, коли створюється тест не з однієї конкретної навчальної дисципліни і не всі завдання в ньому пов'язані з певною галуззю знань. Для перевірки гіпотези про одномірність тесту розроблено багато різних статистичних методів, серед яких найчастіше використовується аналіз власних значень кореляційної матриці, тест Стаута та аналіз залишків одномірної моделі.

Усі моделі латентних характеристик повинні дозволяти порівнювати учасників тестування, розмістивши їх на одній шкалі,

навіть якщо вони виконували не одні і ті ж завдання тесту. Такі вимірювання називають *вільними* (не залежними) від тесту.

1.2. Модель Раша для дихотомічних завдань

Розглянемо деякі передумови введення математичної моделі для тестів з дихотомічними завданнями, коли відповідь на питання оцінюється за допомогою двох позицій: 1 – правильно, 0 – не правильно. Природно вважати, що успіх учасника тестування у вирішенні певного тестового завдання залежить, в основному, від двох чинників: складності завдання та рівня підготовленості випробовуваного. В загальному випадку можна говорити лише про ймовірність успіху у вирішенні конкретного питання і кількісно вимірювати його деяким числом p з відрізка [0; 1]. Отже припускаємо, що ймовірність того, що певний учасник тестування правильно вирішить певне завдання, є деякою функцією двох аргументів – рівня підготовленості випробовуваного s і рівня складності цього завдання t:

$$p = p(s, t) \tag{1.1}$$

Таку функцію називають *функцією успіху*. Тут s і t – поки лише просто символи, що позначають певні терміни.

Припустимо, що в деякому розумінні складність t_1 деякого завдання в k раз більша за складність t_2 іншого завдання $t_1 = k \cdot t_2$, а рівень підготовленості s_1 одного опитаного в k раз більший за рівень підготовленості s_2 іншого. Тобто $\frac{s_1}{t_1} = \frac{s_2}{t_2}$. У такій ситуації природно припустити, що ймовірність правильної відповіді першим учасником (більш підготовленим) на перше завдання (більш важке) повинна збігатися з ймовірністю правильної відповіді другим учасником (менш підготовленим) на друге завдання (менш важке). З цього випливає, що:

1) поняття s і t тісно пов'язані між собою, і не можна визначити одне з них, не визначивши значення іншого;

2) функція (1.1) є однорідною 1-го порядку, тобто імовірність успіху залежить не від кожного аргументу s і t окремо, а лише від їх відношення:

$$p = p(s,t) = p_1(\zeta), \quad \zeta = \frac{s}{t} \tag{1.2}$$

Змінні s і t є *латентними* (не спостережуваними) параметрами, оскільки вони описують деякі приховані характеристики учасників тестування і тестових завдань. Для їх кількісного зіставлення серед багатьох учасників і багатьох завдань необхідно домовитися про відповідні одиниці вимірювання. Для цієї мети простіше всього одному із завдань тесту – за певними ознаками стандартному для даної області – приписати одиничну складність $t_0 = 1$. Тоді про складність t будь-якого іншого завдання тесту можна говорити, що це завдання в t раз важче (якщо $t > 1$) або легше (якщо $0 < t < 1$) стандартного, одиничного. Аналогічні міркування справедливі і щодо рівнів підготовленості s учасників тестування.

Тому значення змінних s і t (а, отже, і $\zeta = s/t$) зручно обирати з числового проміжку $(0;+\infty)$. Це область визначення функції (1.2), а множина її можливих значень – відрізок [0; 1], $p \in [0;1]$.

Навіть не маючи аналітичного виразу функції $p_1(\zeta)$, природно на неї накласти такі умови:

3) функція $p_1(\zeta)$ повинна бути гладкою (неперервною разом зі своєю похідною) і монотонно зростаючою на всій області визначення, оскільки будь-яке збільшення відношення s/t повинне приводити до збільшення імовірності правильної відповіді на питання;

4) $\lim\limits_{\zeta \to 0} p_1(\zeta) = p_1(0) = 0$, що робить безнадійним успіх абсолютно непідготовленого учасника тестування;

5) $\lim\limits_{\zeta \to +\infty} p_1(\zeta) = 1$, що гарантує успіх учаснику тестування, рівень підготовленості якого у багато разів перевищує складність завдання;

6) $p_1(1) = 0{,}5$, тобто максимальна невизначеність в прогнозі результату виконання завдання повинна бути у тому випадку, коли рівень підготовленості учасника збігається зі складністю завдання $t = s$.

Усі ці властивості має S – подібна функція, зображена на рис.1, якщо точкою перегину буде точка (1; 0.5).

Найпростіша аналітична модель ймовірності успіху (1.2), що є однорідною функцією 1-го порядку, запропонована данським математиком Г. Рашем (G.Rasch) у кінці 50-х років минулого століття і має вигляд

$$p = p(s,t) = \frac{s}{s+t} = \frac{s/t}{1+s/t} = \frac{\zeta}{1+\zeta} = p_1(\zeta) \qquad (1.3)$$

Легко перевірити, що ця функція задовольняє всі перераховані вище вимоги 1) – 6). Крім того, функція (1.3) дозволяє наочно інтерпретувати процес виконання завдання складності t учасником тестування з рівнем підготовленості s за допомогою класичної моделі теорії ймовірності – діставання з урни різнокольорових куль. Дійсно, нехай в урні є t чорних куль і s білих. Тоді ймовірність правильного виконання завдання складності t учасником тестування з рівнем підготовленості s збігається із ймовірністю того, що навмання витягнута з урни одна куля виявиться білою. Ця ймовірність і визначається формулою (1.3), де ζ – відношення кількості білих і чорних куль в урні.

Аргументи функції (1.3) є латентними параметрами – їх не можна виміряти безпосередньо. Проте значення самої функції – ймовірність p (або її оцінка) – доступні для вимірювання при реальному тестуванні. У цьому сенсі цікавою є функція

$$\zeta = p_1^{-1}(p), \qquad (1.4)$$

обернена до ймовірності успіху. Вона обов'язково існує, оскільки функція $p = p_1(\zeta)$ монотонна. Її графік симетричний до графіка функції (1.3) відносно бісектриси 1-го квадранта. Оскільки (1.4) виражає латентний параметр ζ через величину p, доступну в реальному тестуванні для вимірювань, то її називають *функцією вимірювання.* З виразу (1.3) легко одержати, що функція вимірювання має вигляд

$$\zeta = \frac{p}{1-p} = \frac{p}{q}, \text{ де } q = 1 - p \qquad (1.5)$$

Таким чином, багатократні повторні випробування, що дозволяють статистично оцінити імовірність успіху, дозволяють оцінити з допомогою (1.5) і латентний параметр ζ. Наприклад, в моделі з повторною вибіркою куль потрібно підрахувати відносну частоту появи білої кулі (при багатократному витягуванні однієї

кулі навмання і поверненні її назад в урну) і скористатися формулою (1.5).

На практиці аргументи $s \in (0;\infty)$ і $t \in (0;\infty)$ часто зручно виражати в логарифмічному масштабі. Введемо для цього наступні позначення

$$\ln s = \theta,\ \ln t = \delta \Leftrightarrow s = e^{\theta},\ t = e^{\delta} \tag{1.6}$$

Тоді функція успіху (1.3) набуває вигляду

$$p = \frac{e^{\theta}}{e^{\theta} + e^{\delta}} = \frac{e^{\theta-\delta}}{1+e^{\theta-\delta}} = \frac{1}{1+e^{-(\theta-\delta)}} = \frac{1}{1+\exp[-(\theta-\delta)]} = \psi(\theta-\delta) \tag{1.7}$$

і називається ***основною логістичною моделлю Раша***. Тут p – ймовірність того, що учасник тестування з рівнем підготовленості θ правильно виконає завдання складності δ. Ймовірність успіху залежить, по суті, тільки від одного параметра – різниці $\theta - \delta$, і тому модель (1.7) називається *однопараметричною*. Припускається, що аргументи $\theta \in (-\infty;\infty)$ і $\delta \in (-\infty;\infty)$ вимірюються однією і тією ж шкалою, одиницю вимірювання назвемо *логіт*. При цьому очевидні співвідношення:

$$\lim_{(\theta-\delta)\to+\infty} p = 1, \quad \lim_{(\theta-\delta)\to-\infty} p = 0 \tag{1.8}$$

і

$$p = 0.5,\ \text{якщо}\ \theta = \delta \tag{1.9}$$

Різниця параметрів $(\theta - \delta)$ має цікаву геометричну інтерпретацію, якщо значення параметра θ розглядати як положення випробовуваного, а значення δ – як положення завдання на одній і тій же осі. Тоді абсолютна величина різниці $|\theta - \delta|$ – це відстань, на якій знаходиться опитаний з рівнем підготовки θ, від завдання зі складністю δ. Якщо ця різниця велика по модулю і негативна, то завдання даремне для вимірювання рівня знань даного студента. Студент напевно не може виконати його вірно, рівень його знань значно нижчий, ніж складність запропонованого завдання. Великі позитивні значення цієї різниці свідчать про те, що опитаний легко впорається з таким завданням, бо його рівень підготовки набагато більший, ніж складність завдання. З точки зору підходу, запропонованого в IRT, такі завдання неефективні для оцінювання

даного значення θ.

Звичайно, в тому випадку, коли θ трохи більше δ, випробуваний може помилитися в завданні, хоча, швидше за все, виконає його вірно. При від'ємних значеннях різниці $(\theta-\delta)$ випробуваного, найімовірніше, чекає невдача, крім виняткових ситуацій, коли можливо вгадування правильної відповіді.

Якщо θ_i – рівень підготовки *i*-го учасника тестування $(i=1,2,\ldots,n)$, а δ_j – складність *j*-го завдання $(j=1,2,\ldots,k)$, то за формулою (1.7) отримаємо ймовірність p_{ij} того, що *i*-ий учасник вірно відповість на *j*-те запитання (тобто, отримає 1 бал):

$$p_{ij}=P\{x_{ij}=1\,|\,\theta_i,\delta_j\}=\left(1+\exp\left[-(\theta_i-\delta_j)\right]\right)^{-1}, \qquad (1.10)$$

де $x_{ij}=\begin{cases}1, & \text{якщо } i-\text{ий опитаний відповів вірно на } j-\text{те завдання};\\ 0, & \text{якщо } i-\text{й опитаний відповів невірно на } j-\text{те завдання.}\end{cases}$

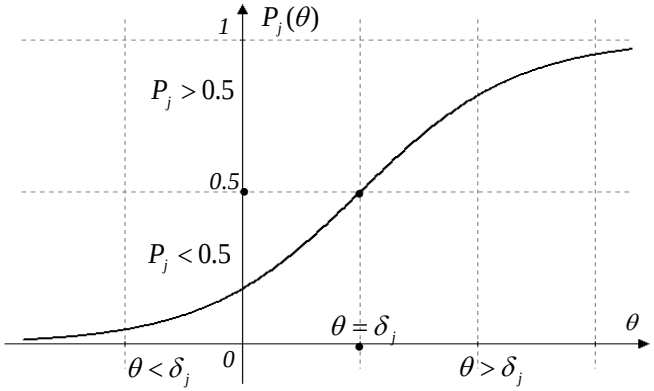


Рис. 2.

Якщо в (1.7) зафіксувати $\delta=\delta_j$, то умовна ймовірність

$$P_j(\theta)=P\{x_{ij}=1\,|\,\delta_j\}=\left(1+\exp[-(\theta-\delta_j)]\right)^{-1} \qquad (1.11)$$

як функція θ повністю характеризує можливості учасників тестування з різними рівнями підготовленості θ при вирішенні завдання складності δ_j і називається *характеристичною функцією за-*

вдання складності δ_j. Її графік (рис.2) називають *характеристичною кривою j-го завдання* (ICC – item characteristic curve).

Для учасника тестування, рівень підготовки якого збігається зі складністю завдання $\theta = \delta_j$, ймовірність правильної відповіді на дане питання становить 0.5. Для учасників із більшим рівнем підготовки ймовірність правильної відповіді на дане питання монотонно зростає. Така властивість характеристичної функції завдання легко інтерпретується і узгоджується з практичним досвідом педагога.

Нахил кривої (крутизна) у кожній точці θ визначається значенням першої похідної:

$$\frac{dP_j(\theta)}{d\theta} = \frac{\exp\left(-(\theta-\delta_j)\right)}{\left(1+\exp\left(-(\theta-\delta_j)\right)\right)^2} = P_j(\theta)\cdot\left(1-P_j(\theta)\right) = P_j(\theta)\cdot Q_j(\theta) \qquad (1.12)$$

Тут $Q_j(\theta) = 1 - P_j(\theta)$ – ймовірність неправильної відповіді на *j*-те запитання. Крива практично горизонтальна (похідна дорівнює 0) при $P = 0$ та $P = 1$. Для цих ділянок кривої різниця між ймовірностями правильної відповіді для двох різних учасників, навіть з великою різницею у рівнях підготовки, незначна.

Знайдемо другу похідну:

$$\frac{d^2P_j(\theta)}{d\theta^2} = \frac{\exp\left(-(\theta-\delta_j)\right)\cdot\left(1-\exp\left(-(\theta-\delta_j)\right)\right)}{\left(1+\exp\left(-(\theta-\delta_j)\right)\right)^3} = P_j(\theta)\cdot Q_j(\theta)\cdot\left(1-2P_j(\theta)\right).$$

Крутизна кривої буде найбільшою (тангенс кута нахилу дотичної до осі θ дорівнює 0.25) при $P = 0.5$ у точці $\theta = \delta_j$, оскільки перша похідна тут має максимум, а сама точка $\theta = \delta_j$ є точкою перегину кривої. Для двох учасників, які мають рівні підготовки з околу точки $\theta = \delta_j$, різниця між ймовірностями правильних відповідей вже буде значною.

Таким чином, маємо важливу властивість: завдання з певним рівнем складності δ_j найкраще диференціює (відрізняє) тих учасників тестування, які мають рівень підготовки $\theta = \delta_j$.

Збільшення складності *j*-го завдання тесту на константу *c* $(c > 0)$ викличе зсув характеристичної кривої вправо. Оскільки

$\theta - \delta_j = (\theta + c) - (\delta_j + c)$, то учасник тестування з більшим рівнем підготовки $\theta + c$ на складніше питання складності $\delta_j + c$ відповідатиме з такою ж ймовірністю, як менш підготовлений учасник на менш складне питання, тобто значення функції $P_j(\theta)$ не зміниться. Це дає підставу для висновку про відносну інваріантність (незалежність) рівня підготовки опитаних від складності завдань тесту. Характеристичні криві різних завдань у моделі Раша не перетинаються і мають однаковий нахил дотичної у точках перетину з прямою $P = 0.5$. На рис.3 зображені характеристичні криві трьох завдань зі складностями $\delta_1 = -1.5$, $\delta_2 = 0.25$ та $\delta_3 = 2.0$.

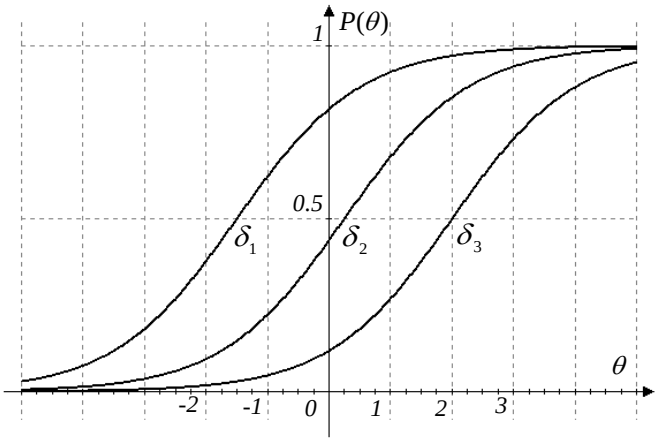


Рис. 3.

Якщо в (1.7) зафіксувати $\theta = \theta_i$, то умовна ймовірність

$$P_i(\delta) = P\{x_{ij} = 1 \mid \theta_i\} = (1 + \exp[-(\theta_i - \delta)])^{-1} \qquad (1.13)$$

як функція δ повністю описує потенційні можливості індивідуума з рівнем підготовленості θ_i при виконанні різних за складністю δ завдань і тому називається *характеристичною функцією рівня підготовленості* θ_i. ЇЇ графік (рис. 4) ще називають *індивідуальною кривою і-го опитаного* (PCC – personal characteristic curve).

Аналогічно отримаємо, що даний учасник тестування вико-

нає вірно завдання складності $\delta = \theta_i$ з ймовірністю 0.5. Чим складніше завдання ($\delta > \theta_i$), тим менша ймовірність його виконання

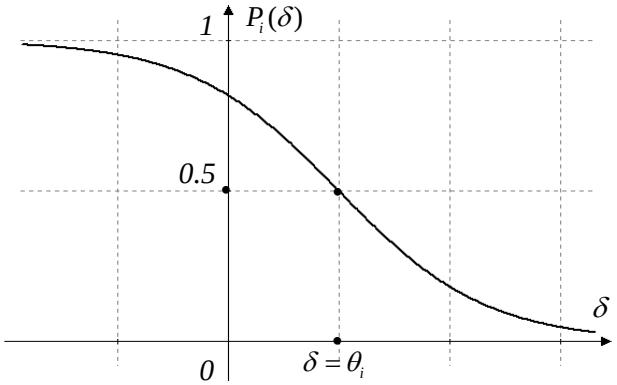


Рис. 4.

даним учасником. Значення змінної $\delta = \theta_i$ є розв'язком рівняння $\frac{d^2 P_i(\delta)}{d\delta^2} = 0$, точкою перегину кривої є точка $(\theta_i; 0.5)$.

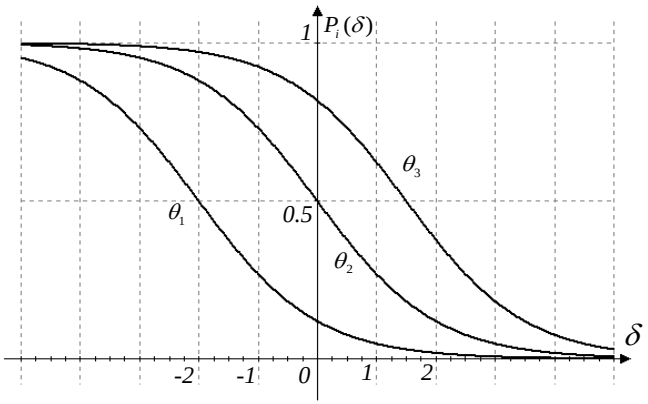


Рис. 5.

Характеристична функція рівня підготовленості θ_i отримується із стандартної характеристичної кривої рівня підготовленості 0 зсувом (без деформації) вздовж осі абсцис. Це означає збільшен-

ня (якщо $\theta_i > 0$) або зменшення (якщо $\theta_i < 0$) ймовірності успішного виконання особою з $\theta = \theta_i$ завдань будь-яких рівнів складності. Отже, характеристичні криві, відповідні різним рівням підготовленості, не перетинаються.

На рис.5 зображені індивідуальні криві трьох учасників тестування з різними рівнями підготовки $\theta_1 = -2.0$, $\theta_2 = 0$ та $\theta_3 = 1.5$.

В теорії IRT функції (1.11) та (1.13) ще називаються Item та Person Response Functions (IRF та PRF).

Зауваження 1. Ймовірність правильної відповіді в моделі Раша залежить лише від різниці $(\theta_i - \delta_j)$, де одиницю вимірювання параметрів θ_i та δ_j назвали *логітом*. З (1.10) отримуємо, що

$$\theta_i - \delta_j = \ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \ln\left(\frac{p_{ij}}{q_{ij}}\right). \qquad (1.14)$$

Отже, *логіт рівня підготовки* – це натуральний логарифм відношення шансів на вірну відповідь на питання «нульової» складності. Аналогічно, *логіт рівня складності* – відношення шансів провалу для особи «нульового» рівня підготовки.

Таблиця 1.

Рівень підготовки θ_i	**Складність завдання** δ_j	**Різниця** $\theta_i - \delta_j$	**Ймовірність вірної відповіді** p_{ij}	**Інформація у питанні** $p_{ij} \cdot q_{ij}$
5	0	5	0,99	0,01
4	0	4	0,98	0,02
3	0	3	0,95	0,05
2	0	2	0,88	0,11
1	0	1	0,73	0,20
0	0	0	0,50	0,25
0	1	-1	0,27	0,20
0	2	-2	0,12	0,11
0	3	-3	0,05	0,05
0	4	-4	0,02	0,02
0	5	-5	0,01	0,01

У таблиці 1 наведено співвідношення між рівнями підготовки θ_i, складностями завдань δ_j, різницями $\theta_i - \delta_j$ у логітах та відповідною їм ймовірністю (1.10). Верхні шість рядків показують ймовірність правильної відповіді осіб з різними рівнями підготовки на питання нульової складності. Шість нижніх рядків показують ймовірність правильної відповіді для особи з нульовим рівнем підготовки на питання різних рівнів складності. Дані цієї таблиці ніяк не зміняться, якщо до величин θ і δ додати будь-яку константу. Це означає, що початок відліку на шкалі логітів можна вибирати довільно (найчастіше це середній рівень підготовки усіх опитаних або середня складність усіх завдань).

Теоретично параметри можуть змінюватись від $-\infty$ до $+\infty$, але на практиці достатньо розглядати інтервал $[-5;5]$, оскільки за його межами ймовірність практично стабільна.

Останній стовпчик таблиці 1 містить добутки $p_{ij}(1-p_{ij})$, які можна трактувати як кількість інформації про різницю $\theta_i - \delta_j$, яка міститься у відповідному елементі матриці відповідей. Корисно відзначити, що інформативність відповідей залежить тільки від відстані $|\theta_i - \delta_j|$ і помітно зменшується із збільшенням цієї відстані. Так, одне завдання максимальної ефективності рівносильне (з точки зору підтримки однієї і тієї ж точності вимірювання) біля 25 завдань мінімальної ефективності. Схожий за змістом результат ми отримали аналітично.

Отже, шкала, у якій вимірюються латентні параметри, є *інтервальною* або *метричною*. Її характерною рисою є наявність метрики та відсутність початку відліку. Така шкала підходить для фіксації взаємного положення вимірюваних об'єктів (один щодо іншого), але вона не в змозі вказати місцезнаходження об'єкта в деякій єдиній системі координат (відстань від початку відліку). З математичної точки зору така ситуація означає, що на множині визначена метрика (введено поняття "відстань"), але немає поняття норми (не визначено поняття "довжина").

Зауваження 2. Підтвердити висновок про відносну інваріантність (незалежність) рівня підготовки опитаних від складності завдань тесту можна, порівнюючи шанси на успіх для двох учас-

ників з різними рівнями підготовки θ_i та θ_k. Відповідно до (1.10) шанси на успіх для цих осіб мають вигляд:

$$\frac{p_{ij}}{1-p_{ij}} = e^{(\theta_i - \delta_j)} \quad \text{та} \quad \frac{p_{kj}}{1-p_{kj}} = e^{(\theta_k - \delta_j)} .$$

Відношення шансів на успіх для *i*-ої та *k*-ої особи дорівнює $e^{(\theta_i - \theta_k)}$. Воно залежить лише від рівнів підготовки цих осіб, і не залежить від складності завдання, яке вони виконували.

Зауваження 3. Функція успіху (1.7) належить відомій в теорії ймовірностей множині логістичних функцій розподілу ймовірностей вигляду $F(x) = (1+e^{-dx})^{-1}$ з дисперсією $D\{x\} = \frac{1}{3}\left(\frac{\pi}{d}\right)^2$, де *d*– параметр. Точка перегину всіх відповідних кривих має координати (0; 0.5), але кривизна кривих в точці перегину пропорційна величині параметра *d*, оскільки $F'(0) = \frac{d}{4}$. Отже, в моделі Раша маємо $d = 1$, нахил усіх кривих в точці перегину однаковий.

Зауваження 4. Часто можна зустріти роботи, у яких логістична функція (1.7), функції (1.11) та (1.13) мають множник 1.7:

$$P_j(\theta) = \left(1+\exp[-1.7(\theta-\delta_j)]\right)^{-1} \quad \text{та} \quad P_i(\delta) = \left(1+\exp[-1.7(\theta_i-\delta)]\right)^{-1}.$$

Цей множник використовується для сумісності логістичної моделі Раша з моделлю нормальної огіви, у якій імовірність правильної відповіді на питання виражається інтегралом нормального розподілу, що дозволяє використовувати замість логістичних кривих добре відому інтегральну функцію стандартного нормального розподілу

$$F(\theta-\delta_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\theta-\delta_j} e^{-\frac{1}{2}x^2} dx .$$

Встановлено, що для одних і тих же значень *x* ординати точок графіків функцій $F(x)$ та $\psi(1.7x)$ відрізняються один від одного досить мало: $|F(x) - \psi(1.7x)| < 0.01$. Найбільш сильний аргумент на користь логістичної функції пов'язаний не з якістю вимірювань, а з відносною простотою її аналітичного задання, вигідною при оцінюванні параметрів θ і δ. Тому в практичних додатках перевагу

зазвичай віддають функції $\psi(1{,}7x)$, яка належить також до сім'ї однопараметричних моделей.

1.3. Моделі IRT для дихотомічних завдань

Логістичну модель Раша інколи називають однопараметричною моделлю сучасної теорії тестів IRT. Однак послідовники вимірювань за моделлю Раша категорично проти такого приєднання, оскільки логіка розвитку моделей суттєво різна, хоча вони і є математично еквівалентними. ***Однопараметричною моделлю IRT***, яку позначають 1PL (One-Parameter Logistik Model), будемо називати модель, згідно з якою характеристична крива завдання визначається співвідношенням:

$$P_j(\theta) = P\left\{x_{ij} = 1 \middle| \delta_j\right\} = \frac{\exp d(\theta - \delta_j)}{1 + \exp d(\theta - \delta_j)} = \left[1 + \exp(-d(\theta - \delta_j))\right]^{-1}. \quad (1.15)$$

Сталий множник d вказує на те, що всі криві мають однаковий нахил при перетині з прямою $p = 0.5$, який дорівнює $0.25 \cdot d$, тобто всі завдання мають однакову роздільну здатність. На оцінки рівня підготовленості та складності завдань множник d не впливає. При $d = 1$ отримуємо модель Раша. 1PL модель має всі властивості моделі Раша.

Якщо тест містить завдання з різною диференціюючою (роздільною) здатністю, то однопараметрична модель 1PL вже не може описати такі емпіричні дані. Для подолання такої проблеми А.Бірнбаум (A.Birnbaum) ввів ще один параметр – *диференціюючу здатність завдання* d_j. Він запропонував умовну ймовірність правильного виконання j-го завдання, яке має два параметри: складність δ_j та диференціюючу здатність d_j, записати у вигляді

$$P_j(\theta) = P\left\{x_{ij} = 1 \middle| \delta_j, d_j\right\} = \left[1 + \exp(-D \cdot d_j(\theta - \delta_j))\right]^{-1}. \quad (1.16)$$

Це ***двопараметрична логістична модель 2PL*** (Two-Parameter Logistik Model). Тут сталий множник $D = 1.7$ для кращого узгодження з моделлю нормальної огіви (зауваження 4). Цей факт дозволяє різні обчислення з логістичною моделлю інтерпретувати, при необхідності, з позиції детально вивченого нормально-

го закону розподілу ймовірностей. Якщо $D \cdot d_j$ для всіх завдань однакове, маємо 1PL модель, якщо $D \cdot d_j = 1$ для всіх завдань – модель Раша. У рамках 2PL моделі кожному тестовому завданню певної складності δ_j може відповідати кілька кривих з різними кутами нахилу дотичної (рис.6), які перетинаються в єдиній точці перегину $(\theta = \delta_j;\ 0.5)$. В цій точці кривизна кожної кривої даної множини, тобто величина тангенса кута нахилу дотичної, дорівнює $Dd_j \cdot 0.25$ (легко переконатись, аналізуючи першу та другу похідні функції (1.16)).

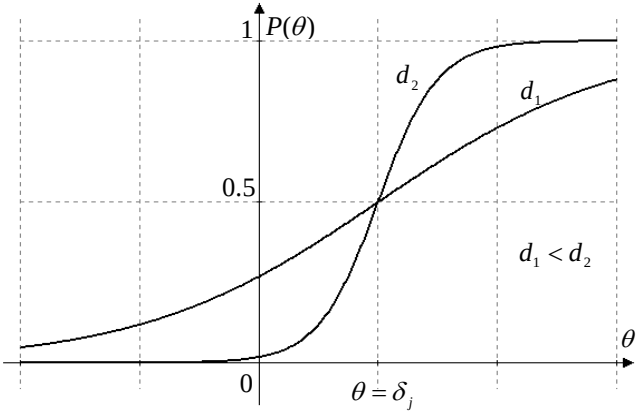


Рис. 6.

Привабливішим є те завдання, параметр d_j якого більший. При малих d_j характеристична крива трудності δ_j є похилою. На практиці це означає, що учасники тестування з хорошим та поганим рівнем підготовки виконують дане завдання з приблизно рівним успіхом. Такі завдання не дають ніякої інформації про індивідуальні відмінності опитаних. Навпаки, якщо d_j велике, то шанси успішного виконання даного завдання учасниками з $\theta < \delta_j$ і $\theta > \delta_j$ істотно відрізняються. Аналізуючи завдання однакової складності, можна зменшити довжину тесту, якщо відібрати серед них кращі завдання з вищою диференціюючою здатністю. Теоре-

тично параметр d_j може змінюватися в інтервалі $(-\infty; +\infty)$, але завдання з від'ємними значеннями d_j не повинні включатися у тест, оскільки тупий кут нахилу дотичної призводить до того, що учасники з вищим півнем підготовки мають менші шанси на правильну відповідь. На практиці, як правило, рекомендують залишати у тесті завдання з d_j, які знаходяться в інтервалі $(0.5; 2.5)$.

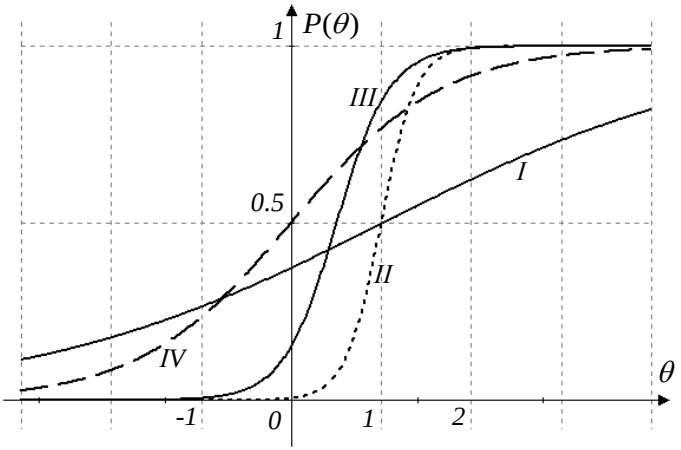


Рис. 7.

Двопараметрична модель має один важливий недолік, який утруднює аналіз тесту: характеристичні криві можуть перетинатися у багатьох точках. Одне і те ж завдання може бути для однієї групи опитаних легшим за інше, а для іншої групи – навпаки, що не завжди має обґрунтоване пояснення. З огляду на це прихильники моделі Раша не визнають процес двопараметричного моделювання за справжнє вимірювання.

На рис.7 зображено характеристичні криві чотирьох завдань різних рівнів складності та різної дискримінуючої здатності: $\delta_1 = \delta_2 = 1.0$, $\delta_3 = 0.5$, $\delta_4 = 0.0$ та $d_1 = 0.3$, $d_2 = 3.0$, $d_3 = 2.0$, $d_4 = 0.7$. У деяких випадках для факту перетину характеристичних кривих можна знайти природне пояснення. Наприклад, завдання I з поганою дискримінуючою здатністю для слабких опитаних може виявитися легшим, ніж для сильних через можливість угадування правильної відповіді. Щоб отримати математичну мо-

дель, яка б більш точно описувала емпіричні дані, А.Бірнбаум запропонував врахувати ефект угадування за допомогою ще одного параметра.

Нехай c_j – ймовірність того, що опитаний з рівнем підготовки θ вгадає правильну відповідь на j-те питання. У процесі опитування може відбутися дві події: або правильна відповідь вгадана, або відповідь не вгадується і тоді ймовірність правильної відповіді на питання знаходиться за формулою (1.16). Остаточно ймовірність правильної відповіді особи з рівнем підготовки θ на j-те запитання з трьома параметрами δ_j, d_j, c_j можна записати за формулою повної ймовірності:

$$P_j(\theta) = P\left\{x_{ij} = 1 \middle| \delta_j, d_j, c_j\right\} = c_j + \frac{(1-c_j)}{1+\exp(-D\cdot d_j(\theta-\delta_j))}. \qquad (1.17)$$

Це і є ***трипараметрична модель 3PL*** (Three-Parameter Logistik Model) з *параметром угадування* c_j. Характеристична функція (1.17) вже не є логістичною. При $\theta \to -\infty$ границя функції дорівнює c_j (для всіх попередніх моделей така границя дорівнювала нулеві), тому графік функції (1.17) має горизонтальну асимптоту $p = c_j$ (рис.8). Це означає, що навіть дуже слабо підготовлені учасники тестування мають не нижчу за параметр угадування c_j ймовірність вірно відповісти на питання.

Точкою перегину характеристичної кривої завдання є точка з абсцисою $\theta = \delta_j$, як у попередніх моделях. Але значення функції в цій точці більше за 0.5. Ймовірність вірної відповіді для учасників з області, де завдання найбільш інформативне, збільшується. При цьому зменшується дискримінуюча здатність завдання, крива більш пологіша, оскільки з першої похідної у точці $\theta = \delta_j$ випливає, що $tg\alpha = (Dd_j - Dd_j c_j)\cdot 0.25$.

Початковою оцінкою параметра угадування можна вибирати величину, обернено пропорційну до кількості варіантів відповідей у завданнях з вибором. Наприклад, якщо у завданні з однією правильною відповіддю пропонується чотири варіанти відповіді, то

ймовірність вгадати її $c_j = 0.25$. У процесі аналізу це значення уточнюється, але не бажано, щоб воно було більшим.

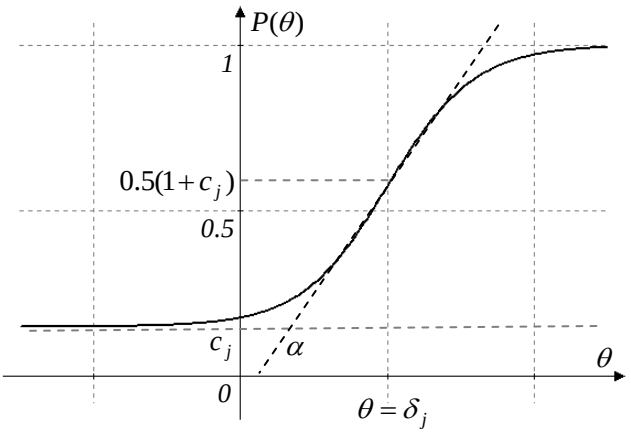


Рис. 8.

Недоліком даної моделі є також те, що характеристичні криві різних завдань можуть перетинатися, що значно ускладнює аналіз і обробку даних у процесі вдосконалення тесту (рис.9).

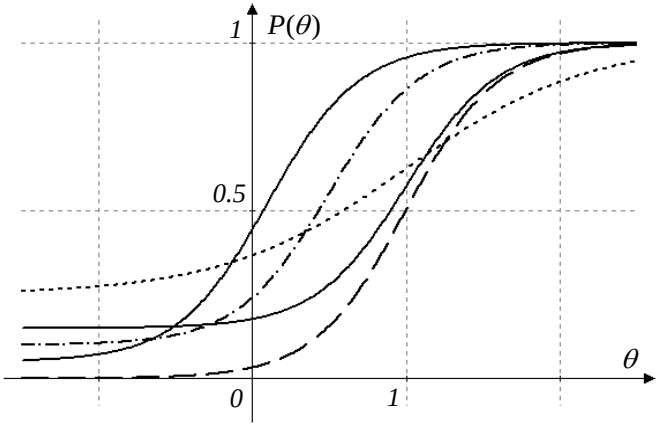


Рис. 9.

Двопараметрична модель отримується з трипараметричної,

якщо $c_j = 0$, а однопараметрична при $c_j = 0$ та $Dd_j = Const$. Отже, моделі 2PL та 3PL введені з метою вдосконалення, щоб математична модель якомога точніше описувала емпіричні дані. Ідеологія ж вимірювань у моделі Раша є зовсім іншою: якщо модель не відповідає емпіричним даним, то вдосконалюватися повинні саме дані. Дехто вважає, що лише модель Раша відповідає вимогам до якісного вимірювального інструменту.

Можна ввести і більшу кількість параметрів, що більш повно характеризують тестові завдання і учасників тестування. Проте точність їх оцінок за результатами масового тестування при цьому помітно знижується, погіршується збіжність ітераційних методів, що використовуються у процесі оцінювання.

Зауваження 1. Шкала, у якій вимірюються латентні характеристики, має довільний початок та одиницю вимірювання. Як правило, їх вибирають так, щоб середнє значення оцінок латентної характеристики дорівнювало нулеві, а стандартне відхилення – одиниці для деякої контрольної групи опитаних. Тоді отримуються криві на деякому симетричному відносно точки 0 проміжку.

Значення параметрів d_j та δ_j залежать від одиниці вимірювання і шкали, вибраної для θ. Зміна одиниці вимірювання та нульової точки шкали рівносильне лінійному перетворенню $\theta' = k\theta + m$, де k та m – довільні дійсні числа. При перетворенні θ в θ' величини d_j та δ_j теж зміняться:

$$P_j(\theta') = \frac{1}{1+\exp\left(-D\cdot d_j\left(\frac{\theta'-m}{k}-\delta_j\right)\right)} = \frac{1}{1+\exp\left(-D\cdot\frac{d_j}{k}\left(\theta'-(k\delta_j+m)\right)\right)}.$$

Отже, у новій системі координат $d_j' = \frac{d_j}{k}$, а $\delta_j' = k\delta_j + m$. Ймовірність $P_j(\theta)$ при цьому не змінюється, оскільки виконується рівність $d_j'(\theta' - \delta_j') = d_j(\theta - \delta_j)$. Параметр c_j при перетворенні шкали не змінюється.

1.4. Математичні моделі для політомічних завдань

Дихотомічне завдання допускає лише дві можливі категорії відповіді: вірно/невірно або 1/0. Політомічне завдання – більше двох категорій, наприклад: вірно/частково вірно/невірно або 2/1/0.

Можна виділити чотири основних джерела політомічності (багатоваріантності) тестових завдань, коли результат представляється у вигляді деяких впорядкованих категорій. Перше з них – повторні випробування, коли опитані мають фіксовану незалежну кількість спроб відповісти на кожне питання. Спостережене значення успіхів x може набувати значень від 0 до кількості спроб m. Такий формат корисний при дослідженні психомоторних навичок, тоді x – кількість спроб, у яких завдання виконано правильно. У даному випадку порядок появи успіхів не враховується, а лише кількість. Впорядковані категорії відповідей тут мають значення 0, 1, 2,. . ., m.

Другий тип впорядкованих категорій виникає тоді, коли немає верхньої межі кількості незалежних успіхів або невдач. Спостереженим значенням x може бути, наприклад, кількість успішно завершених спроб за указаний проміжок часу, кількість помилок, які робить людина при читанні заданого абзацу тощо. Впорядковані категорії відповідей тут мають значення 0, 1, 2,..., ∞. Математичні моделі для аналізу таких завдань були запропоновані Рашем у 1960-х роках та досліджені його послідовниками.

Третій тип політомічності походить від рейтингових шкал, коли відповіді на кожне завдання можуть оцінюватись фіксованим набором альтернатив, наприклад: погано, посередньо, добре. Або коли потрібно висловити своє відношення до кожного пункту анкети за допомогою альтернатив типу не підтримую, підтримую частково, підтримую повністю. Особливістю такого формату даних є те, що для кожного завдання повинен використовуватись один і той самий набір категорій 0, 1, 2,. . . , m.

Четвертий тип впорядкованих категорій відповідей може виникати тоді, коли завдання потребує покрокового виконання і кожен крок оцінюється як частковий успіх певною кількістю балів. Мотивом для такого оцінювання є надія, що це дасть більш точну характеристику здібностей опитаних, ніж оцінка типу здав – не здав. Для кожного j-го завдання може використовуватись свій на-

бір категорій 0, 1, 2. . . . m_j. До даного типу політомічності можна віднести і рейтингові шкали (типу Лайкерта). Тоді перехід від однієї позиції шкали до іншої можна вважати одним кроком.

Найбільш вживаною математичною моделлю для таких завдань є модель Partial Credit Model (PCM), яку запропонував Дж.Мастерс (G.Masters) у 1982 р. Дослівний переклад «Модель часткових кредитів» вживається рідко. Інколи модель PCM називають політомічною моделлю Раша.

Нехай за виконання j-го завдання тесту ($j=\overline{1,k}$) i-ий опитуваний ($i=\overline{1,n}$) може одержати x_{ij} балів, де $x_{ij}=0,1,\ldots,m_j$. Щоб дістати найвищу категорію (рівень) m_j, учасник повинен послідовно подолати m_j кроків: на першому кроці спочатку потрібно досягти першого рівня в один бал, потім на другому кроці досягти категорії в два бали і так далі. У таблиці 2 наведено різні можливі

Таблиця 2.

Учасник, i	Рівень виконання				Бали, x_{ij}
	0 y_{0ij}	**1** y_{1ij}	**2** y_{2ij}	**3** y_{3ij}	
	Перший крок	Другий крок	Третій крок		
1	1 →	1 →	1 →	1	3
2	1				0
3	1 →	1 →	1		2
4	1 →	1			1
5	1				0
6	1 →	1 →	1 →	1	3
7	1 →	1			1
8	1 →	1 →	1		2
9	1				0
10	1 →	1 →	1 →	1	3
Всього	$s_{j0}=10$	$s_{j1}=7$	$s_{j2}=5$	$s_{j3}=3$	

ситуації для 10 учасників при відповіді на трикрокове завдання. Дихотомічна змінна y_{lij} приймає значення 1, якщо *i*-ий учасник подолав *l*-ий крок *j*-го завдання, і 0 – якщо ні. У даному випадку

сім учасників подолали перший крок і досягли рівня 1, п'ять учасників подолали другий крок і досягли рівня 2, а три учасники повністю впоралися із завданням і досягли найвищого рівня 3. При цьому жоден учасник не може досягти *l*-го рівня, якщо до цього він не досяг (*l-1)*-го рівня.

Рівень складності виконання *l*-го кроку *j*-го завдання позначимо через δ_{jl} $\left(l=\overline{0,m_j}\right)$, причому вважаємо, що цей параметр не несе ніякої інформації про складність попередніх кроків. Тобто, це не складність досягнення *l*-го кроку, як у деяких інших моделях, а складність переходу з одного кроку на інший. Таким чином, умовна ймовірність того, що буде подолано *l*-ий крок (дихотомічна змінна $y_{lij}=1$), якщо перед цим було подолано (*l-1)*-ий крок, залежить лише від рівня підготовки учасника тестування θ_i та від складності переходу між кроками δ_{jl} і повинна зростати на всій області визначення латентної змінної рівня підготовленості. Це дає підставу застосувати на цьому кроці логістичну модель Раша (1.10):

$$p_{lij}^{ум}=P\left\{y_{lij}=1\,\middle|\,\theta_i,\delta_{jl}\right\}=\left(1+\exp\left[-(\theta_i-\delta_{jl})\right]\right)^{-1}.$$

З іншого боку, ця умовна ймовірність за означенням

$$p_{lij}^{ум}=\frac{p_{lij}}{p_{(l-1)ij}+p_{lij}}, \tag{1.18}$$

де чисельник вказує на ймовірність того, що *i*-та особа у *j*-ому завданні виконає всі *l* кроків (відповідно набере *l* балів, а не (*l -1)*), а знаменник – ймовірність того, що ця особа виконає *l* або (*l-1)* кроків. Аналогічно можна було б трактувати умовну ймовірність для дихотомічних завдань, але там $p_{1ij}^{ум}=p_{1ij}$, оскільки $p_{0ij}+p_{1ij}=1$.

Отже, $\dfrac{p_{lij}}{p_{(l-1)ij}+p_{lij}}=\dfrac{1}{1+\exp(-(\theta_i-\delta_{jl}))}$. Звідси отримуємо рекурентне співвідношення

$$p_{lij}=p_{(l-1)ij}\cdot e^{(\theta_i-\delta_{jl})}. \tag{1.19}$$

Послідовно можна виразити усі безумовні ймовірності через p_{0ij} :

$$
\begin{aligned}
p_{1ij} &= p_{0ij} \cdot e^{(\theta_i - \delta_{j1})}, \\
p_{2ij} &= p_{1ij} \cdot e^{(\theta_i - \delta_{j2})} = p_{0ij} \cdot e^{(\theta_i - \delta_{j1}) + (\theta_i - \delta_{j2})}, \\
p_{3ij} &= p_{2ij} \cdot e^{(\theta_i - \delta_{j3})} = p_{0ij} \cdot e^{(\theta_i - \delta_{j1}) + (\theta_i - \delta_{j2}) + (\theta_i - \delta_{j3})}, \qquad (1.20) \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
p_{m_j ij} &= p_{(m_j-1)ij} \cdot e^{(\theta_i - \delta_{jm_j})} = p_{0ij} \cdot e^{(\theta_i - \delta_{j1}) + (\theta_i - \delta_{j2}) + (\theta_i - \delta_{j3}) + \ldots + (\theta_i - \delta_{jm_j})}.
\end{aligned}
$$

Оскільки кожен учасник наберe хоча б якусь кількість балів від 0 до m_j, то сума всіх ймовірностей дорівнює 1:

$$p_{0ij} + p_{1ij} + p_{2ij} + p_{3ij} + \ldots + p_{m_j ij} = 1. \qquad (1.21)$$

Підставляючи в (1.21) вирази з (1.20), отримаємо

$$p_{0ij}\left[1 + e^{(\theta_i - \delta_{j1})} + e^{(\theta_i - \delta_{j1}) + (\theta_i - \delta_{j2})} + \ldots + e^{(\theta_i - \delta_{j1}) + (\theta_i - \delta_{j2}) + (\theta_i - \delta_{j3}) + \ldots + (\theta_i - \delta_{jm_j})}\right] = 1$$

Для зручності далі в усіх формулах вважатимемо $e^{(\theta_i - \delta_{j0})} = 1$. Тоді

$$p_{0ij} = \frac{1}{\sum_{h=0}^{m_j} \exp \sum_{g=0}^{h} (\theta_i - \delta_{jg})}. \qquad (1.22)$$

Повертаючись до рівностей (1.20), можемо записати загальну формулу для ймовірності того, що *i*-та особа у *j*-ому завданні виконає рівно *l* кроків (або ж отримає *l* балів):

$$p_{lij} = P\left\{x_{ij} = l \,\middle|\, \theta_i, (\delta_j)\right\} = \frac{\exp \sum_{g=0}^{l} (\theta_i - \delta_{jg})}{\sum_{h=0}^{m_j} \exp \sum_{g=0}^{h} (\theta_i - \delta_{jg})}, \ (l = 0,1,2,\ldots,m_j) \qquad (1.23)$$

Формула (1.23) і визначає модель ***Partial Credit***. Тут (δ_j) – вектор складностей усіх кроків *j*-ого завдання. Зауважимо, що у чисельнику фігурують складності лише тих l кроків, які дана особа подолає, а у знаменнику – складності всіх можливих у даному завданні кроків.

На рис.10 та 11 показана залежність умовних (верхня части-

на рисунка) та безумовних (нижня частина) ймовірностей від рівня підготовки опитаних θ для деякого двокрокового завдання. Якщо складність другого кроку вища, ніж першого $\delta_{j2} > \delta_{j1}$ (рис.10), то ймовірність подолати два кроки, а не один, на всій множині рівнів підготовки θ є меншою.

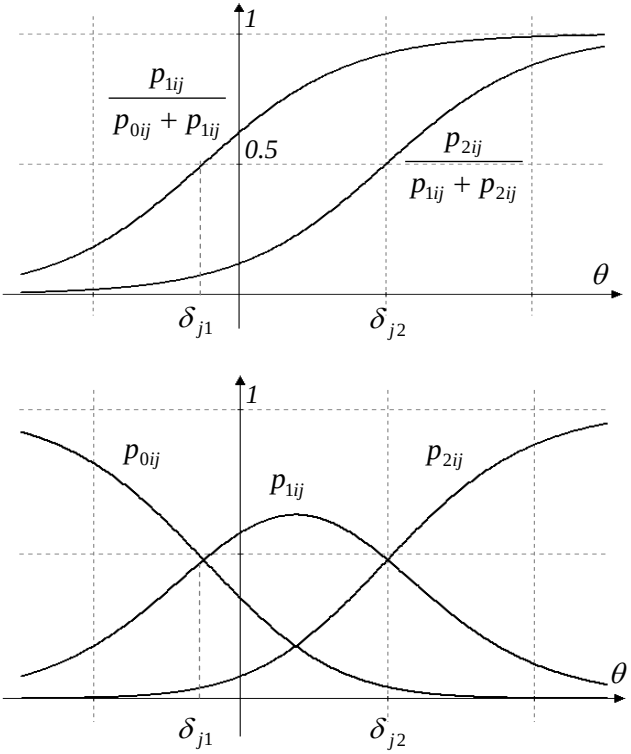


Рис. 10.

Складність кожного кроку δ_{jl} є абсцисою точки перетину кривих безумовних ймовірностей двох сусідніх категорій (*l-1)* та *l*, тобто, це те значення рівня підготовки, де ймовірності досягти кожної з даних категорій однакові: $p_{(l-1)ij} = p_{lij}$. Ординати точок перетину не обов'язково дорівнюють 0.5, на відміну від деяких інших моделей.

Для учасників тестування з рівнями підготовки θ між δ_{j1} та δ_{j2} найбільш ймовірною є можливість подолати саме перший крок. Якщо складності обох кроків стають близькими, для кожного учасника тестування ймовірність завершити лише перший крок зменшується, а збільшується ймовірність завершити обидва кроки, або жодного. Якщо ж складність другого кроку менша, ніж першого $\delta_{j2} < \delta_{j1}$ (рис.11), то ще менш ймовірно для кожного учасника завершити лише один крок. Таке завдання перетворюється на дихотомічне, де кожен учасник має більше шансів отримати 0 або 2 бали. Категорія в 1 бал не працює, її не доцільно виділяти у даному завданні.

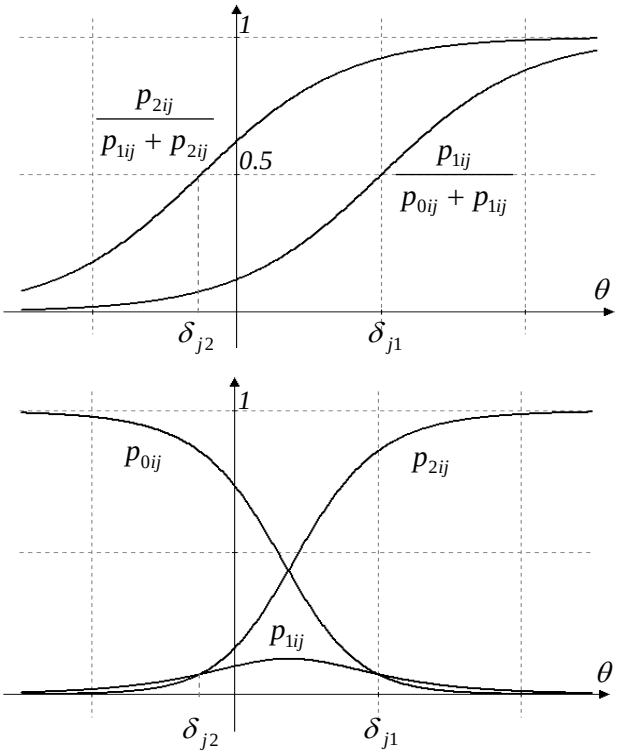


Рис. 11.

Функції, зображені на рис.10 та 11, мають назву Item Category Response Function (ICRF), аналогічно до Item Response Function (IRF) у дихотомічному випадку. На одному рисунку зображувати та аналізувати ICRF для кількох різних завдань не зовсім зручно через велику кількість кривих та точок їх перетину. Але можна порівнювати *характеристичні криві* ICC різних завдань, які у загальному випадку вказують на залежність очікуваного бала (математичного сподівання) за завдання від рівня підготовки. У випадку політомічних завдань характеристична крива завдання ICC визначається рівнянням:

$$E_j(\theta) = \sum_{l=0}^{m_j} l \cdot p_{lij} = 0 \cdot p_{0ij} + 1 \cdot p_{1ij} + 2 \cdot p_{2ij} + \ldots + m_j \cdot p_{m_j ij}. \qquad (1.24)$$

Для дихотомічних завдань криві ICC та IRC дійсно збігаються, оскільки очікуваний бал $E_j(\theta) = 0 \cdot p_{0ij} + 1 \cdot p_{1ij} = P_j(\theta)$.

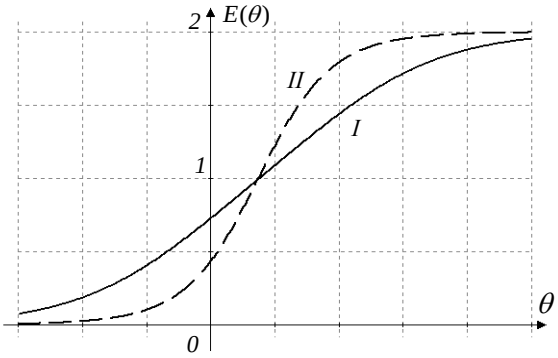


Рис. 12.

На рис.12 зображено ICC двох двокрокових завдань. Перше завдання (суцільна лінія) має складності кроків $\delta_{11} = -0.5, \delta_{12} = 2.0$ (як на рис.10), для другого завдання (пунктирна лінія) навпаки, $\delta_{21} = 2.0, \delta_{22} = -0.5$ (як на рис.11).

Зауваження. У моделі Partial Credit кожне завдання може мати різну кількість категорій відповідей, а кожна ICC має свою асимптоту, що відповідає максимальній категорії для даного за-

вдання. Криві ICC для різних завдань можуть мати точки перетину, що ускладнює аналіз тесту.

Ще до появи моделі Partial Credit у 1978 р. Д.Андріч (D.Andrich) запропонував математичну модель Rating Scale Model (RSM) для аналізу анкет з впорядкованими відповідями, у якій важливо, щоб усі запитання мали однакову кількість категорій $(l = 0,1,2,\ldots,m)$. Зараз ця модель отримується з моделі Partial Credit як частинний випадок.

Якщо у моделі PCM не накладається ніяких обмежень на складність кроків для різних завдань, деякі кроки можуть бути легші в одному завданні і складніші в іншому, то у моделі RSM складність кроків всередині кожного завдання не повинна змінюватись від завдання до завдання. Щоб це врахувати, будемо вважати, що на рівень складності виконання *l*-го кроку *j*-го завдання у моделі Partial Credit впливають два параметри – складність самого завдання та складність кроку, яка однакова для всіх завдань, тобто

$$\delta_{jl} = \delta_j + \tau_l . \qquad (1.25)$$

На рис.13 показано, як формуються складності кроків у RSM.

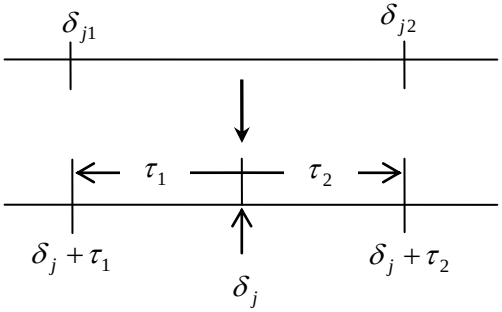


Рис. 13.

Враховуючи (1.25) у основній формулі (1.23) для моделі PCM, отримаємо

$$p_{lij} = \frac{\exp\sum_{g=0}^{l}\left(\theta_i - (\delta_j + \tau_g)\right)}{\sum_{h=0}^{m}\exp\sum_{g=0}^{h}\left(\theta_i - (\delta_j + \tau_g)\right)}, \quad (l = 0,1,2,\ldots,m).$$

Перегрупувавши доданки, отримаємо відому раніше основну

формулу моделі ***Rating Scale***:

$$p_{lij} = \frac{\exp\left[-\sum_{g=0}^{l}\tau_g + l\cdot(\theta_i - \delta_j)\right]}{\sum_{h=0}^{m}\exp\left[-\sum_{g=0}^{h}\tau_g + h\cdot(\theta_i - \delta_j)\right]}, (l = 0,1,2,\ldots,m) \quad (1.26)$$

Графіки CBRF для всіх завдань мають однакову форму, але їх положення на горизонтальній осі визначається складністю δ_j кожного завдання, криві ICC для різних завдань не перетинаються (як у моделі Раша).

Моделі Partial Credit та Rating Scale належать до сім'ї моделей Раша, оскільки володіють основними властивостями моделі Раша. Зауважимо, що у обох моделях не враховується дискримінуюча здатність завдання (вона однакова для всіх завдань).

При побудові математичних моделей для політомічних завдань спочатку переважав підхід, коли впорядковані категорії відповідей розглядались як граничні точки (пороги), що розділяють деяку неперервну область. Значення порогів залежать від складності досягнення відповідного рівня.

Популярну модель на основі такого підходу Graded Response Model (GRM) запропонувала Ф.Самеїма (F.Samejima) ще у 1969 р. В основі моделі такі міркування: коли особа *i* отримує питання *j* , у якому є можливість подолати *m* категорій, з'являється деяка латентна випадкова змінна ε_{ij} . Ймовірність того, що ця змінна набуде значення вище, ніж *l*-та категорія, залежить від рівня підготовки особи θ_i , складності подолання даного порогу λ_{jl} , $(l = 1,2,\ldots,m)$ та диференціюючої здатності завдання a_j . Якщо припустити, що ε_{ij} має логістичний розподіл, який схожий на нормальний, але з важчими хвостами та більшим ексцесом, (див. зауваження 3 пункту 1.2), то ймовірність того, що *i*-та особа у *j*-му завданні подолає *l*-ий або вищий поріг матиме вигляд:

$$p^*_{lij} = \frac{\exp\left(a_j(\theta_i - \lambda_{jl})\right)}{1+\exp\left(a_j(\theta_i - \lambda_{jl})\right)} = \left(1+\exp\left[-a_j(\theta_i - \lambda_{jl})\right]\right)^{-1}. \quad (1.27)$$

У даній моделі λ_{jl} – складність подолання *l*-го порогу, тоді як у попередніх моделях δ_{jl} – складність переходу між сусідніми порогами, причому тут $\lambda_{j1} < \lambda_{j2} < \ldots < \lambda_{jm}$. Вище значення подоланого порогу завжди відповідає вищому рівню підготовки. На рис.14 показано зв'язок між ймовірностями (1.27) та ймовірностями p_{lij} того, що особа у даному завданні подолає рівно *l* порогів (на прикладі завдання з двома порогами).

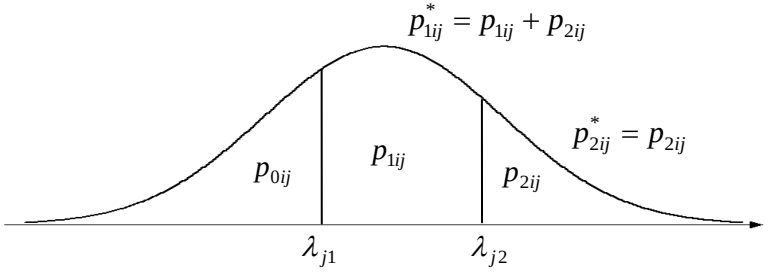


Рис. 14.

Ймовірності p_{lij} можна знайти послідовним відніманням (для даного прикладу):

$$p_{0ij} = 1 - p^*_{1ij} = \frac{1}{\Psi}\left(1 + \exp\left[a_j(\theta_i - \lambda_{j2})\right]\right),$$

$$p_{1ij} = p^*_{1ij} - p^*_{2ij} = \frac{1}{\Psi}\left(\exp\left[a_j(\theta_i - \lambda_{j1})\right] - \exp\left[a_j(\theta_i - \lambda_{j2})\right]\right),$$

$$p_{2ij} = p^*_{2ij} = \frac{1}{\Psi}\left(\exp\left[a_j(\theta_i - \lambda_{j2})\right] + \exp\left[a_j(2\theta_i - \lambda_{j1} - \lambda_{j2})\right]\right),$$

де $\Psi = \left(1 + \exp\left[a_j(\theta_i - \lambda_{j1})\right]\right) \cdot \left(1 + \exp\left[a_j(\theta_i - \lambda_{j2})\right]\right)$.

У загальному ж випадку отримаємо основну формулу моделі ***Graded Response***:

$$p_{lij} = P\left\{x_{ij} = l \,\middle|\, \theta_i, a_j, (\lambda_j)\right\} = p^*_{lij} - p^*_{(l+1)ij} = \qquad (1.28)$$

$$= \frac{1}{1+\exp\left[-a_j(\theta_i - \lambda_{jl})\right]} - \frac{1}{1+\exp\left[-a_j(\theta_i - \lambda_{j(l+1)})\right]}.$$

Застосовуючи формулу (1.28) для $l = \overline{0, m}$, потрібно мати на увазі, що $p^*_{0ij} = 1$ та $p^*_{(m+1)ij} = 0$. Це має очевидний ймовірнісний зміст: кожна особа подолає *0*-поріг або вище і жодна – (*m+1*)-поріг.

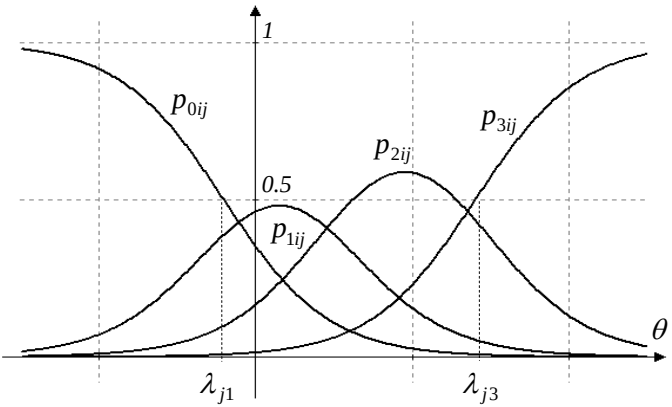


Рис. 15.

На рис.15 представлені функції ICRF для деякого завдання з чотирма категоріями (трьома порогами). Тут, як і у попередніх двох моделях, монотонними є лише ICRF, що відповідають двом крайнім категоріям. Два крайніх порогових значення λ_{j1} та λ_{j3} відповідають рівням підготовки, коли $p_{0ij} = 0.5$ та $p_{3ij} = 0.5$. Якщо відстань між λ_{j1} та λ_{j3} збільшується, то ймовірності проміжних категорій p_{1ij} та p_{2ij} зростають. Якщо складності крайніх порогів дуже близькі, ймовірності проміжних категорій зменшуються і завдання перетворюється на дихотомічне 0/3. Інші значення порогових складностей не мають якоїсь інтерпретації на графіках функцій ICRF.

У межах одного завдання коефіцієнт дискримінації a_j не змінюється, це дозволяє уникнути зайвих перетинів кривих та

від'ємних ймовірностей. Але навіть якщо $a_j = 1$, модель GRM не має властивостей, притаманних моделям сімейства Раша. Зокрема, не вдається алгебраїчно відокремити латентні параметри завдань та учасників тестування.

На рис. 16 проілюстровано два основні підходи до дихотомізації кроків при побудові математичних моделей для тестів із завданнями політомичного типу. У рамках цих підходів, крім розглянутих вище моделей, пропонуються і інші вдосконалені або гібридні моделі.

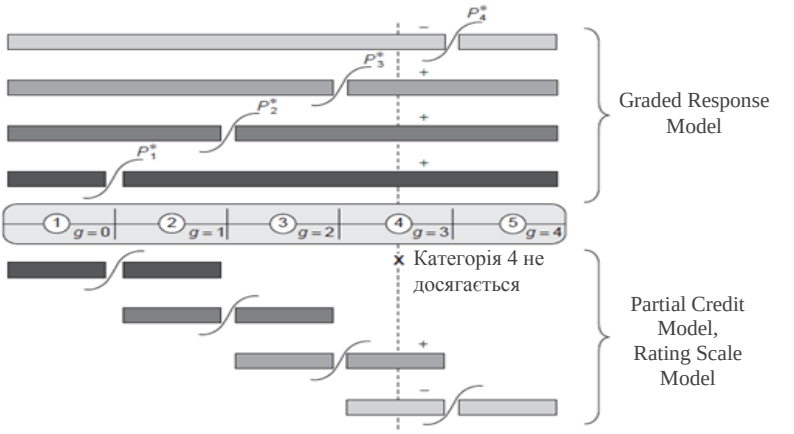


Рис. 16.

Р.Бок (R.Bock) у 1972 р. запропонував для аналізу політомічних завдань модель Nominal Response Model (NRM), альтернативну до попередніх моделей, у якій нема заздалегідь визначеної залежності між порядком взаємовиключних категорій та рівнями підготовки. Тобто, не накладається умова, що вибір вищої категорії відповідає вищому рівню підготовки.

Ймовірність того, що особа з рівнем підготовки θ_i обере категорію l у j-ому завданні $(l = 1,2,\ldots,m)$ у моделі ***Nominal Response*** має вигляд:

$$p_{lij} = \frac{\exp\left(a_{jl}\theta_i + c_{jl}\right)}{\sum_{k=1}^{m} \exp\left(a_{jk}\theta_i + c_{jk}\right)}, \qquad (1.29)$$

де a_{jl} - параметри диференціації, які змінюються для різних ICRF навіть у межах одного завдання, c_{jl} - інтерсепти кривих ICRF, які відображають популярність категорії. Величину $z_{jl} = a_{jl}\theta_i + c_{jl}$ називають багатовимірним логітом, а сама модель будується за принципом багатовимірної моделі Раша. Додатково вводяться обмеження на параметри:

$$\sum_{k=1}^{m} a_{jk} = \sum_{k=1}^{m} c_{jk} = 0 .$$

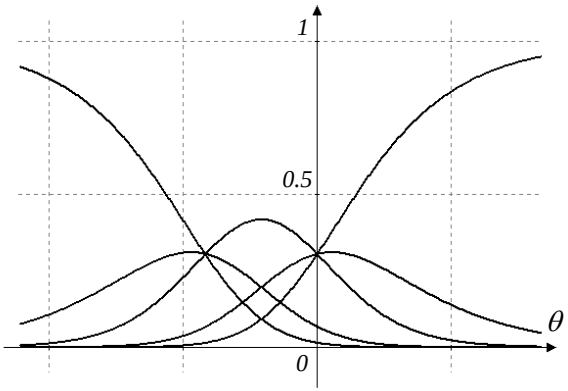


Рис.17.

Строго монотонними будуть лише ICRF, які відповідають найменш та найбільш популярним категоріям. На рис.17 зображено ICRF для питання з п'ятьма категоріями. NRM дозволяє визначити, який порядок варіантів відповідей асоціюється з високим рівнем латентної змінної підготовленості опитаних. Ця модель використовується також для визначення положення нейтральної відповіді серед впорядкованих відповідей на шкалі Лайкерта.

Всі розглянуті у даному розділі одновимірні математичні моделі сучасної теорії тестів представлено у Таблиці 3. Зірочкою позначені моделі, які відносять до моделей сімейства Раша. Більш детально деякі властивості цих моделей будуть розглянуті у наступних розділах.

У даному посібнику не розглядаються моделі багатовимірної сучасної теорії тестів (Multidimensional Item Response Theory (MIRT)), компонентні моделі (Linea Logistic Test Model (LLTM), General Latent Trait Model (GLTM)), непараметричні моделі (Monotone Homogeneity Model (MHM), Double Monotonicity Model (DMM)) та багато інших моделей, інформацію про які можна знайти в енциклопедії з освітніх вимірювань [14] та в іншій довідковій літературі.

Таблиця 3.

Модель	**Тип завдань**	**Особливості моделі**
Раша*/ однопараметрична 1PL	дихотомічні	Дискримінуюча здатність для всіх завдань стала.
Двопараметрична 2PL	дихотомічні	Враховується дискримінуюча здатність завдань.
Трипараметрична 3PL	дихотомічні	Враховується дискримінуюча здатність завдань та угадування.
Graded Response Model	політомічні	Впорядковані категорії, враховується дискримінуюча здатність завдань.
Nominal Response Model	політомічні	Не впорядковані категорії, враховується дискримінуюча здатність завдань.
Partial Credit Model *	політомічні	Дискримінуюча здатність для всіх завдань стала, складності кроків можуть змінюватись.
Rating Scale Model *	політомічні	Дискримінуюча здатність для всіх завдань стала, складності кроків для всіх завдань однакові.

2. ОЦІНЮВАННЯ ЛАТЕНТНИХ ПАРАМЕТРІВ

2.1. Властивості первинних балів

Нехай у тестуванні беруть участь n учасників різного рівня підготовленості θ_i, ($i = 1,2,\ldots,n$). Кожному учаснику пропонується один і той же варіант тесту, що складається з k завдань різної складності δ_j, ($j = 1,2,\ldots,k$). Результат виконання кожного завдання оцінюється за дихотомічним принципом: ставиться одиниця, якщо завдання виконано правильно, і нуль, якщо завдання виконано невірно. Множина всіх таких одиниць і нулів утворює прямокутну таблицю – матрицю розмірності $n \times k$. Позначатимемо цю матрицю буквою $A = (a_{ij})$ і будемо називати *матрицею відповідей*. Вона має n рядків і k стовпців. Її елемент a_{ij}, що стоїть на перетині i – го рядка і j – го стовпця, виражає можливий результат виконання i – м учасником j – го завдання. Елементи a_{ij} є величинами випадковими: вони приймають значення 1 з ймовірністю $p_{ij} = p(\theta_i, \delta_j)$, яка у даному випадку має вигляд (1.10), і значення 0 з імовірністю $q_{ij} = 1 - p_{ij}$. При потребі розрізняти випадкові величини від їх реалізації будемо позначати останні як $\tilde{a}_{ij}$.

Матриця відповідей є тією початковою інформацією, за якою передбачається оцінити латентні параметри тестування – складності завдань δ_j та рівні підготовленості випробовуваних θ_i. Матриця має яскраво виражену вертикальну структуру – кількість стовпців k (число завдань в одному тесті) звичайно не перевищує 60 – 70, але кількість рядків n (число учасників тестування) має порядок, як правило, декількох сотень або декількох тисяч. Таким чином, матриця відповідей містить декілька десятків або навіть сотень тисяч нулів і одиниць. Проте такий величезний масив початкових даних часто вдається істотно зменшити (редукувати) без втрати інформації. В основі такої редукції лежить поняття достатньої статистики. Взагалі *статистикою* називається будь-яка функція результатів початкових спостережень (в нашому випадку елементів a_{ij} матриці відповідей A), що проявляє статистичну стій-

кість в тому сенсі, що значення цієї функції при повторних спостереженнях може бути передбачено з достатньо кращою точністю, ніж результат окремого спостереження. Підкреслимо, що будь-які статистики, будучи функціями початкових спостережень, тобто величин випадкових, також є випадковими. Але розкидання їх можливих значень при повторних спостереженнях значно менше розкидання можливих значень окремих спостережень.

Зручними статистиками при обробці результатів масового тестування є так звані маргінальні (часткові) суми елементів матриці відповідей A по кожному рядку і по кожному стовпцю, тобто числа:

$$b_i \equiv a_{i\bullet} = \sum_{j=1}^{k} a_{ij},\ i = 1,2,...,n \ ; \tag{2.1}$$

$$c_j \equiv a_{\bullet j} = \sum_{i=1}^{n} a_{ij},\ j = 1,2,...,k \tag{2.2}$$

Тут число $b_i = a_{i\bullet}$ – кількість вірно виконаних завдань учасником з номером i , називається *первинним балом і – го учасника.* Воно відображає деяку міру успіху i–го випробовуваного при виконанні k завдань даного тесту.

Число $c_j = a_{\bullet j}$ – кількість учасників, які вірно виконали завдання з номером j. Називатимемо його, по аналогії з попереднім поняттям, *первинним балом j – го завдання*. Різниця $\left(n - c_j\right)$ відображає деяку міру складності j –го завдання при виконанні цього завдання даним континентом n учасників тестування.

Первинні бали, як будь-які інші статистики, є величинами випадковими, але при повторних тестуваннях вони набагато більш стійкі, ніж випадкові елементи матриці відповідей, тобто результати вирішення якого-небудь одного завдання яким-небудь одним учасником.

Випадкові елементи a_{ij} матриці відповідей A в результаті випробування можуть приймати два значення 0 та 1, тому математичне сподівання a_{ij} та дисперсія мають вигляд:

$$M\{a_{ij}\} = 1 \cdot p_{ij} + 0 \cdot q_{ij} = p_{ij}\ , \tag{2.3}$$

$$D\{a_{ij}\} = M\{a_{ij}^2\} - M^2\{a_{ij}\} = 1^2 \cdot p_{ij} + 0^2 \cdot q_{ij} - p_{ij}^2 = p_{ij}q_{ij}. \quad (2.4)$$

Далі легко знайти математичні сподівання і дисперсії первинних балів:

$$M\{b_i\} = M\{\sum_{j=1}^{k} a_{ij}\} = \sum_{j=1}^{k} M\{a_{ij}\} = \sum_{j=1}^{k} p_{ij}\ , \quad (2.5)$$

$$D\{b_i\} = D\{\sum_{j=1}^{k} a_{ij}\} = \sum_{j=1}^{k} D\{a_{ij}\} = \sum_{j=1}^{k} p_{ij}q_{ij}. \quad (2.6)$$

Первинні бали мають *узагальнений біноміальний* розподіл. Розподіл відповідної ймовірності $P_k(b_i)$ зручно описувати за допомогою так званої твірної функції вигляду [3]:

$$\varphi_{ik}(x) = (p_{i1}x + q_{i1})(p_{i2}x + q_{i2}) \cdot \ldots \cdot (p_{ik}x + q_{ik}). \quad (2.7)$$

Ймовірність $P_{ik}(a_{i\bullet})$ того, що учасник з номером *i* із *k* завдань вірно виконає b_i штук, $0 \le b_i \le k$, дорівнює коефіцієнту при x^{b_i} у розкладі функції (2.7) за степенями x.

Наприклад, якщо $k = 2$, то

$$\varphi_{i2}(x) = (p_{i1}x + q_{i1})(p_{i2}x + q_{i2}) = p_{i1}p_{i2}x^2 + (p_{i1}q_{i2} + p_{i2}q_{i1})x + q_{i1}q_{i2}$$

Імовірність $P_{i2}(2)$ того, що *i* –ий учасник вирішить обидва завдання вірно, дорівнює $p_{i1} \cdot p_{i2}$, коефіцієнт при *x* дає імовірність $P_{i2}(1)$ того, що вірно вирішено лише одне будь-яке завдання; коефіцієнт при x^0, тобто вільний член $q_{i1} \cdot q_{i2}$, дорівнює ймовірності $P_{i2}(0)$ того, що обидва завдання вирішено неправильно. Сума коефіцієнтів при всіх можливих степенях *x* дорівнює одиниці, тобто у даному прикладі $P_{i2}(2) + P_{i2}(1) + P_{i2}(0) = 1$.

Ймовірності того, що за наявності *k* завдань первинний бал учасника з номером *i* виявиться: а) меншим за *m*; б) більшим за *m*; в) не меншим за *m*; г) не більшим за *m*, – можна знайти за формулами, відповідно:

$$P_{ik}(0) + P_{ik}(1) + \ldots + P_{ik}(m-1),$$

$$P_{ik}(m+1) + P_{ik}(m+2) + \ldots + P_{ik}(k),$$

$$P_{ik}(m) + P_{ik}(m+1) + \ldots + P_{ik}(k), \qquad (2.8)$$
$$P_{ik}(0) + P_{ik}(1) + \ldots + P_{ik}(m).$$

Якщо припустити, що всі завдання тесту мають однаковий рівень складності, тобто $\delta_j = \delta$ для $\forall j$ від 1 до k та $p_{ij} = p_i$ незалежно від j, то формули (2.5) – (2.7) спрощуються і приймають вигляд:

$$M\{b_i\} = kp; \; D\{b_i\} = kp_i q_i, \qquad (2.9)$$

$$\varphi_{ik}(x) = (p_i x + q_i)^k, \qquad (2.10)$$

де $p_i = p(\theta_i, \delta)$, $q_i = 1 - p_i$. Коефіцієнтами при x^b, $0 \le b \le k$, в розкладі бінома твірної функції (2.10) за степенями x будуть біноміальні коефіцієнти. Отже,

$$P_{ik}(b) = C_k^b \cdot p_i^b \cdot q_i^{k-b}. \qquad (2.11)$$

Співвідношення (2.11) називається *формулою Бернуллі*, а відповідний розподіл ймовірності називається біноміальним. Той факт, що для однаково складних завдань у тесті первинний бал b має біноміальний розподіл, ми записуватимемо у вигляді

$$b \in Bi(k, p)$$

Отже, біноміальний розподіл – це розподіл числа «успіхів» в k незалежних випробуваннях з двома результатами («успіх» - «неуспіх») і сталою ймовірністю «успіху» $p \in (0;1)$. Зокрема

$$a_{ij} \in Bi(k, p_{ij}),$$

де a_{ij} – елемент матриці відповідей, $p_{ij} = p(\theta_i, \delta_j)$ – функція успіху. Зауважимо, що вирази для дисперсії справедливі тільки в припущенні незалежності випадкових величин a_{ij}, а для справедливості формул для математичного сподівання ця незалежність не потрібна.

На практиці формулу (2.11) зручно замінити наближеним співвідношенням (заміна тим точніша, чим більше k) вигляду (локальна теорема Лапласа):

$$P_{ik}(b) \approx \frac{1}{\sqrt{kp_i q_i}} \varphi(x), \qquad (2.12)$$

де $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, $x = \frac{b - kp_i}{\sqrt{kp_i q_i}}$.

Функція $\varphi(x)$ є густиною розподілу ймовірності нормальної центрованої (тобто має нульове математичне сподівання) і нормованої (тобто має одиничну дисперсію) випадкової величини *u*, тобто $u \in N(0;1)$.

Ймовірність попадання первинного бала b_i в заданий проміжок $[m_1, m_2]$ зручно прогнозувати не за допомогою формул (2.8), а за допомогою інтегральної теореми Лапласа:

$$P_{ik}(m_1 \le b_i \le m_2) \approx \Phi(x_2) - \Phi(x_1) \ . \qquad (2.13)$$

Тут $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{u^2}{2}} du$ – функція Лапласа,

$$x_1 = \frac{m_1 - kp_i}{\sqrt{kp_i q_i}}, \quad x_2 = \frac{m_2 - kp_i}{\sqrt{kp_i q_i}}$$

Формули (2.12), (2.13) зручно використовувати для наближеного прогнозу результатів тестування. Для значень функцій $\varphi(x)$ та $\Phi(x)$ використовують статистичні таблиці.

Для прикладу розглянемо тест, який містить *k* = 20 завдань однакової складності, а *p* позначає середню ймовірність правильної відповіді на одне завдання цього тесту. Тоді учасників тестування з різним рівнем підготовленості θ_i можна класифікувати за величиною відповідної ймовірності $p_i = p(\theta_i, \delta)$. Умовно називатимемо рівень підготовленості дуже слабким, якщо $p_i < 0{,}3$; слабким, якщо $0{,}3 \le p_i \le 0{,}55$; середнім, якщо $0{,}35 < p_i \le 0{,}75$; добрим, якщо $p_i > 0{,}75$.

Таблиця 4 містить ймовірності $P_{ik}(j \le b \le 20)$ того, що учасники тестування різного рівня наберуть кількість первинних балів *b* не менше *j* = 1, 2,..., 20, які обчислені за формулами (2.13).

Таблиця 4.

j	**Рівень підготовленості**							
	Дуже слабкий	Слабкий			Середній		Хороший	
	p = 0,2	**p = 0,3**	**p = 0,4**	**p = 0,5**	**p = 0,6**	**p = 0,7**	**p = 0,8**	**p = 0,9**
1	0,99	1,00	1,00	1,00	1,00	1,00	1,00	1,00
2	0,90	0,98	1,00	1,00	1,00	1,00	1,00	1,00
3	0,72	0,95	1,00	1,00	1,00	1,00	1,00	1,00
4	0,50	0,85	0,98	1,00	1,00	1,00	1,00	1,00
5	0,29	0,70	0,92	0,99	1,00	1,00	1,00	1,00
6	0,13	0,50	0,83	0,97	1,00	1,00	1,00	1,00
7	0,05	0,31	0,69	0,92	1,00	1,00	1,00	1,00
8	0,01	0,16	0,50	0,82	0,98	1,00	1,00	1,00
9	0,00	0,07	0,32	0,68	0,92	1,00	1,00	1,00
10	0,00	0,03	0,18	0,50	0,83	0,98	1,00	1,00
11	0,00	0,01	0,09	0,33	0,69	0,95	1,00	1,00
12	0,00	0,00	0,03	0,19	0,50	0,85	0,99	1,00
13	0,00	0,00	0,01	0,09	0,32	0,70	0,95	1,00
14	0,00	0,00	0,00	0,04	0,18	0,50	0,87	1,00
15	0,00	0,00	0,00	0,01	0,09	0,31	0,71	0,99
16	0,00	0,00	0,00	0,00	0,03	0,16	0,50	0,93
17	0,00	0,00	0,00	0,00	0,01	0,07	0,29	0,77
18	0,00	0,00	0,00	0,00	0,00	0,03	0,13	0,50
19	0,00	0,00	0,00	0,00	0,00	0,01	0,05	0,23
20	0,00	0,00	0,00	0,00	0,00	0,00	0,01	0,07

Якщо завдання тесту закритої форми п'ятьма варіантами відповіді, з яких тільки один є вірним, а інші лише правдоподібні, то ймовірність $p = 0,2$ відповідає бездумному вгадуванню відповіді. Проте, як видно з таблиці 4, одержати при цьому первинний бал вище 7–8 практично не можливо. Якщо найменша кількість балів, що відповідає ситуації «зараховано», дорівнює 12, то такий бар'єр, згідно аналізу таблиці 4, можуть подолати з ймовірністю вищою за 0,5 лише учасники, рівень підготовки яких не менше, ніж середній.

Високий бал (18 і вище) зуміють одержати, за прогнозом, тільки добре підготовлені учасники тестування.

Первинні бали в результаті тестування є величинами випадковими. При великій кількості числа k завдань в тесті їх розподіл приблизно нормальний. Тому довірчий інтервал для них можна побудувати за допомогою наступного співвідношення :

$$P_{ik}\{|b - M(b)| \le \varepsilon\} = 2\Phi\left(\frac{\varepsilon}{\sqrt{D(b)}}\right) \tag{2.14}$$

Тут математичне сподівання $M(b)$ і дисперсія $D(b)$ визначаються формулами (2.5), (2.6) або (2.9); ε – деяка додатна стала, що визначає довжину інтервалу 2ε.

Більш точний результат можна отримати за допомогою наступного факту:

$$P_{ik}\{m_1 < M(b_i) < m_2\} = d, \tag{2.15}$$

де

$$\begin{aligned} m_1 &= \frac{k}{k+l^2}\left(b_i + \frac{l^2}{2} - l\sqrt{\frac{b_i(k-b_i)}{k} + \frac{l^2}{2}}\right); \\ m_2 &= \frac{k}{k+l^2}\left(b_i + \frac{l^2}{2} + l\sqrt{\frac{b_i(k-b_i)}{k} + \frac{l^2}{2}}\right); \end{aligned} \tag{2.16}$$

l – значення аргументу функції Лапласа, при якому $2\Phi(l) = d$; d – рівень значущості інтервалу (m_1, m_2).

Розглянемо приклад, коли деякий учасник тестування з номером i при виконанні тесту, що складається з $k = 20$ завдань, набрав $b_i = 16$ балів. Яка імовірність й того, що реальне значення первинного бала знаходиться на відрізку [14; 18] ?

Для відповіді на це питання скористаємося спочатку найпростішою схемою (2.14), замінивши потрібну для цього ймовірність p_i її оцінкою $\hat{p}_i = \frac{16}{20} = 0{,}8$ і узявши $\varepsilon = 2$. Маємо

$$2\Phi\left(\frac{2}{\sqrt{20 \cdot 0{,}8 \cdot 0{,}2}}\right) = 2\Phi(1{,}12) = 0{,}74 = d\ .$$

Тепер поставимо завдання навпаки: який інтервал для

$b_i = 16$ слід очікувати з рівнем значущості d = 0,74? Користуючись формулами (2.16), отримаємо

$$m_1 = \frac{20}{20+1{,}12^2}\left(16+0{,}63-1{,}12\sqrt{\frac{16\cdot 4}{20}+0{,}63}\right)=13{,}6;\ m_2 = 17{,}7.$$

Округливши m_1 і m_2 до цілих значень, одержуємо проміжок [14; 18).

Всі формули даного розділу, починаючи з (2.5), та обидва приклади стосуються первинних балів учасників (2.1). Проте все залишається правильним і стосовно первинних балів завдань (2.2). Достатньо лише у формулах (2.5) - (2.16) замінити b_i на c_j, k на n і p_i на p_j.

Наприклад, у тестуванні, в якому взяли участь n = 100 учасників, первинний бал завдання з номером j виявився рівним $c_j = 80$. Який довірчий інтервал для цього первинного бала слід очікувати з рівнем значущості d = 0,95?

Кінці шуканого інтервалу одержимо за формулами (2.16), які у даному випадку матимуть вигляд:

$$\begin{aligned} m_1 &= \frac{n}{n+l^2}\left(c_j+\frac{l^2}{2}-l\sqrt{\frac{c_j(n-c_j)}{n}+\frac{l^2}{2}}\right);\\ m_2 &= \frac{n}{n+l^2}\left(c_j+\frac{l^2}{2}+l\sqrt{\frac{c_j(n-c_j)}{n}+\frac{l^2}{2}}\right). \end{aligned} \qquad (2.17)$$

Отримаємо $l = 1{,}96 \ \Rightarrow \ m_1 = 71,\ m_2 = 87$.

Зауваження 1. Вище наголошувалося на тому, що формули (2.6), (2.9) для дисперсії первинних балів справедливі лише у тому випадку, коли випадкові величини a_{ij} незалежні, тобто ймовірності вирішення деяким учасником різних завдань не залежать одна від одної. В загальному випадку дисперсія визначається формулою

$$D\{b_i\} = \sum_{j=1}^{k} D\{a_{ij}\} + 2\sum_{j<l} k_{i,jl}\ .$$

де $k_{i,jl}$ – коваріаційний момент випадкових величин a_{ij} і a_{il}:

$$k_{i,jl} = M\{a_{ij} \cdot a_{il}\} - M\{a_{ij}\} \cdot M\{a_{il}\}$$

Величина $a_{ij} \cdot a_{il}$ дорівнює одиниці тільки у тому випадку, коли i –й учасник вирішив і j –те, і l –те завдання. Позначимо імовірність цієї події $p_{i,jl}$. Тоді

$$M\{a_{ij} \cdot a_{il}\} = p_{i,jl}\,, \quad k_{i,jl} = p_{i,jl} - p_{ij} \cdot p_{il}\,.$$

Тому $$D\{b_i\} = \sum_{j=1}^{k} p_{ij} \cdot q_{ij} + 2\sum_{j<l} (p_{i,jl} - p_{ij} \cdot p_{il})\,. \qquad (2.18)$$

Отже, для обчислення дисперсії первинного бала недостатньо знати тільки ймовірність p_{ij} правильної відповіді i –го учасника на j –те завдання, потрібно ще знати ймовірність $p_{i,jl}$ правильної відповіді на кожну пару (j, l) завдань.

Зокрема, якщо $p_{ij} = p_i$ для $\forall j$, а $p_{i,jl}$ не залежить від j та l і дорівнює p_i^*, то формула (2.18) спрощується і набуває вигляду

$$D\{b_i\} = kp_i \cdot q_i + k(k-1)(p_i^* - p_i^2),$$

де p_i^* – ймовірність правильної відповіді i–го учасника на будь-яку пару завдань.

Зауваження 2. На практиці часто зручно замінювати різну ймовірність середньою величиною

$$p_i = \frac{1}{k}\sum_{j=1}^{k} p_{ij}\,. \qquad (2.19)$$

При цьому $$M\{b_i\} = kp_i\,, \quad D\{b_i\} = kp_i\,q_i \qquad (2.20)$$

Переконаємося, що така заміна не тільки практично зручна, але і теоретично виправдана. Дійсно, математичні сподівання первинного бала, обчислені за формулами (2.5) і (2.20) збігаються, оскільки $\sum_{j=1}^{k} p_{ij} = p_i k$. Наближене обчислення дисперсії за формулою (2.20) дає завищений результат у порівнянні з точною формулою (2.6). Дійсно, відповідно до (2.6):

$$D\{b_i\} = \sum_{j=1}^{k} p_{ij}q_{ij} = \sum_{j=1}^{k}(p_{ij} - p_{ij}^2) = \sum_{j=1}^{k} p_{ij} - \sum_{j=1}^{k} p_{ij}^2 = p_i k - \sum_{j=1}^{k} p_{ij}^2$$

Легко перевірити, що умовний мінімум суми $\sum_{j=1}^{k} p_{ij}^2$ за умови $\frac{1}{k}\sum_{j=1}^{k} p_{ij} = p_i = const$ досягається при $p_{i1} = p_{i2} = ... = p_{ik} = p_i$. Тому при заміні ймовірностей p_{ij} (j = 1, 2, ..., k) їх середнім значенням p_i дисперсія первинного бала досягає свого максимуму.

Цей несподіваний результат означає, зокрема, що будь-які відхилення складностей δ_j завдань тесту від їх середнього рівня зменшують величину випадкових коливань первинного бала і, відповідно, зменшують розсіяння можливих оцінок відповідних рівнів підготовленості учасників. Зворотне твердження доводиться аналогічно і полягає в наступному: чим однорідніший за рівнем підготовленості склад учасників тестування, тим більше розсіяння оцінок рівня складності завдань слід очікувати. Вказаний взаємозв'язок між дисперсіями контингенту учасників тестування і множиною тестових завдань є найважливішою властивістю тесту з незалежними завданнями, які виконують незалежні учасники. Видалення впливу цього взаємозв'язку з оцінок латентних параметрів – одна з умов об'єктивного калібрування тестових завдань і шкалування учасників.

Важливою властивістю первинних балів є також те, що в рамках деяких математичних моделей вони є статистиками достатніми.

2.2. Достатні статистики

Достатньою статистикою для сімейства розподілів ймовірності $\{P_\lambda : \lambda \in \Lambda\}$ або для параметра (в загальному випадку, векторного) $\lambda \in \Lambda$ називається така статистика X (векторна випадкова величина), що для будь-якої події S існує варіант умовної ймовірності $P_\lambda(S \mid X = x)$, не залежний від λ. Це еквівалентно вимозі, що умовний розподіл будь-якої іншої статистики Y за умови $X = x$ не залежить від λ.

Статистика X буде достатня для сімейства $\{P_\lambda\}$ лише в

тому випадку, коли відповідна густина розподілу P_λ може бути факторизована (розкладена на множники) у вигляді :

$$p_\lambda(\omega) = g_\lambda(X(\omega)) \cdot h(\omega), \qquad (2.21)$$

де g_λ і h – невід'ємні функції події ω, причому g_λ залежить від λ, але від ω залежить тільки через $X(\omega)$, тоді як h не залежить від λ.

Наприклад, учасникам тестування запропоновано k завдань однакової складності. Рівень підготовленості учасника передбачається оцінювати імовірністю ρ правильного розв'язку кожного завдання. Кожному учаснику в матриці відповідей відповідає один рядок, послідовність незалежних випадкових величин, що приймають значення 1 з невідомою імовірністю ρ і значення 0 з імовірністю 1 — ρ (схема Бернуллі). Імовірність конкретної реалізації a_1, a_2,…, a_k вказаної послідовності визначається теоремою множення ймовірностей і має вигляд

$$p_\rho(a_1, a_2, ..., a_k) = \prod_{j=1}^{k} \rho^{a_j} (1-\rho)^{1-a_j} = \rho^{\sum_{j=1}^{k} a_j} \cdot (1-\rho)^{k-\sum_{j=1}^{k} a_j} \qquad (2.22)$$

Рівність (2.21) виконується, якщо припустити

$$X = \sum_{j=1}^{k} a_j, \quad \lambda = \rho, \quad g_\lambda = p_\rho, \quad h = 1,$$

і тому первинні бали учасників є достатніми статистиками – ймовірність (2.22) залежить від результатів тестування $a_1, \ldots, a_k$ тільки через маргінальну суму $X = \sum_{j=1}^{k} a_j$. Зокрема, емпірична частота $\bar{\rho} = \frac{1}{k} \sum_{j=1}^{k} a_j$ є достатньою оцінкою для шуканої імовірності ρ.

Знання достатньої статистика X дає вичерпний матеріал відносно параметра λ, оскільки будь-які додаткові статистичні дані нічого не додають до тієї інформації про параметр, яка міститься в розподілі X. Перехід від початкового сімейства розподілів до сімейства розподілів достатньої статистики називається *редукцією статистичної задачі.* Суть редукції полягає у зменшенні (часто вельми значному) кількості початкових даних без втрати інформації, що міститься в цих даних.

Переконаємося зараз, що в рамках логістичної моделі Раша первинні бали є достатніми статистиками.

Відповідно до (1.3), ймовірність того, що учасник тестування з номером i правильно розв'яже завдання з номером j а, отже, (i, j) – позицію в матриці відповідей A займе 1, визначається формулою (до введення логарифмічного масштабу)

$$p_{ij} = \frac{s_i}{s_i + t_j} = \frac{s_i / t_j}{1 + s_i / t_j} = \frac{\zeta_{ij}}{1 + \zeta_{ij}} \qquad (2.23)$$

де s_i – рівень підготовленості i -го учасника, t_j – рівень складності j-го завдання.

Ймовірність невірного розв'язку (тобто $a_{ij} = 0$), очевидно,

$$q_{ij} = 1 - p_{ij} = 1 - \frac{s_i / t_j}{1 + s_i / t_j} = \frac{1}{1 + s_i / t_j} = \frac{1}{1 + \zeta_{ij}}.$$

Вирази для p_{ij} і q_{ij} можна записати єдиною формулою

$$\frac{(\zeta_{ij})^{a_{ij}}}{1 + \zeta_{ij}} = \frac{(s_i / t_j)^{a_{ij}}}{1 + s_i / t_j} = \begin{cases} p_{ij}, & \text{якщо } a_{ij} = 1, \\ q_{ij}, & \text{якщо } a_{ij} = 0. \end{cases} \qquad (2.24)$$

Тому ймовірність конкретної реалізації $A = (a_{ij})$ матриці відповідей (тут хвильку над конкретними реалізаціями не пишемо, щоб формули виглядали зручніше), на основі теореми множення ймовірностей, має вигляд:

$$p_{s,t}(A_{n\times k}) = \prod_{i=1}^{n}\prod_{j=1}^{k} \frac{(s_i / t_j)^{a_{ij}}}{1 + s_i / t_j} = \frac{\prod_{i=1}^{n}\prod_{j=1}^{k} s_i^{a_{ij}} \cdot \prod_{i=1}^{n}\prod_{j=1}^{k} t_j^{-a_{ij}}}{\prod_{i=1}^{n}\prod_{j=1}^{k} (1 + s_i / t_i)} =$$

$$= \frac{\prod_{i=1}^{n} s_i^{b_i} \cdot \prod_{j=1}^{k}\prod_{i=1}^{n} t_j^{-a_{ij}}}{\prod_{i=1}^{n}\prod_{j=1}^{k} (1 + s_i / t_i)} = \frac{\prod_{i=1}^{n} s_i^{b_i} \cdot \prod_{j=1}^{k} t_j^{-c_j}}{\prod_{i=1}^{n}\prod_{j=1}^{k} (1 + s_i / t_i)}. \qquad (2.25)$$

Останній вираз показує, що дана імовірність залежить від елементів a_{ij} матриці відповідей A тільки через її маргінальні суми

$b_i = a_{i\bullet}$, $c_j = a_{\bullet j}$ по рядках і по стовпцях. Тому первинні бали учасників тестування і завдань тесту є достатніми статистиками для шуканих параметрів s_i і t_j відповідно (i = 1, 2, ..., n; j = 1, 2, ..., k). Отже, при оцінюванні параметрів можна замість $n \cdot k$ елементів матриці відповідей оперувати лише первинними балами, загальна кількість яких рівна (n + k). При цьому кількість початкових даних зменшується в десятки разів без якої-небудь втрати інформації.

Інший важливий висновок, який випливає з достатності первинних балів, полягає в наступному. Статистичні оцінки рівнів підготовленості всіх учасників тестування, що набрали однакову кількість первинних балів, збігаються, оскільки є функціями рівних достатніх статистик. Це підтверджується і іншим означенням достатніх статистик: умовна ймовірність того, що i-ий учасник тестування отримає конкретний рядок відповідей $a_{i1}, a_{i2}, \ldots, a_{ik}$, за умови, що первинний бал дорівнює b_i, не залежить від рівня підготовленості. Дійсно, на основі теореми множення, ймовірність конкретної реалізації $a_{i1}, a_{i2}, \ldots, a_{ik}$ для i-го учасника:

$$P\left\{(a_{ij}) \middle| \theta_i, (\delta_j)\right\} = \prod_{j=1}^{k}\left[\frac{\exp a_{ij}(\theta_i - \delta_j)}{1+\exp(\theta_i - \delta_j)}\right] = \frac{\exp(b_i\theta_i)}{\Psi_i}\exp\left(-\sum_{j=1}^{k} a_{ij}\delta_j\right),$$

де $\Psi_i = \prod_{j=1}^{k}\left(1+\exp(\theta_i - \delta_j)\right)$. Тут використали модель Раша після введення нових змінних (1.10). Ймовірність того, що i-ий учасник отримає рівно b_i балів знаходимо, додаючи ймовірності усіх можливих конкретних реалізацій із загальною сумою балів b_i:

$$P\left\{b_i \middle| \theta_i, (\delta_j)\right\} = \sum_{(b_i)} \frac{\exp(b_i\theta_i)}{\Psi_i}\exp\left(-\sum_{j=1}^{k} a_{ij}\delta_j\right) =$$

$$= \frac{\exp(b_i\theta_i)}{\Psi_i}\sum_{(b_i)}\exp\left(-\sum_{j=1}^{k} a_{ij}\delta_j\right)$$

Тоді умовна ймовірність отримати конкретний вектор реалізацій $a_{i1}, a_{i2}, \ldots, a_{ik}$, якщо сума балів дорівнює b_i, має вид:

$$P\left\{(a_{ij})\middle| b_i,(\delta_j)\right\}=\frac{P\left\{(a_{ij})\middle|\theta_i,(\delta_j)\right\}}{P\left\{b_i\middle|\theta_i,(\delta_j)\right\}}=\frac{\exp(b_i\theta_i)\exp\left(-\sum\limits_{j=1}^{k}a_{ij}\delta_j\right)}{\exp(b_i\theta_i)\sum\limits_{(b_i)}\exp\left(-\sum\limits_{j=1}^{k}a_{ij}\delta_j\right)}=$$

$$=\frac{\exp\left(-\sum\limits_{j=1}^{k}a_{ij}\delta_j\right)}{\sum\limits_{(b_i)}\exp\left(-\sum\limits_{j=1}^{k}a_{ij}\delta_j\right)}. \qquad (2.26)$$

Умовна ймовірність (2.26) не залежить від рівня підготовки учасника тестування, а лише від складності завдань. Тому всі учасники тестування, які наберуть однакову кількість первинних балів, матимуть рівні оцінки латентних параметрів рівня підготовки.

Отже, не знаючи реальних оцінок цих параметрів, навіть не обговорюючи методи їх обчислення, ми можемо упорядкувати всіх учасників тестування в порядку зростання оцінок їх рівнів підготовленості тільки на підставі первинних балів учасників.

Зауваження 1. Необхідно враховувати те, що йдеться про впорядковування саме статистичних оцінок латентних параметрів, а не їх істинних значень, які залишаються неприступними і розташовуються, можливо, в іншому порядку, ніж їх оцінки. В більшій мірі це відноситься до характеристик учасників тестування, ніж до характеристик завдань, оскільки точність відповідності оцінок параметрів (величин випадкових) їх істинним значенням (константам) залежить від об'єму початкових даних, доступних для обчислення оцінок. Тому проявляється той факт, що первинні бали учасників виводяться на підставі k вихідних елементів матриці відповідей (маргінальні суми по рядках), а при обчисленні первинних балів завдань використовується n таких елементів (маргінальні суми по стовпцях), причому n, як правило, значно перевершує k.

Наприклад, якщо тест, що складається з 20 завдань, виконують 1000 учасників тестування, то $n = 1000$, $k = 20$. Тому первинні бали кожного з 1000 учасників можуть приймати тільки обмежене число (21) дискретних значень 0, 1, 2, ..., 19, 20. В той же час, кіль-

кість можливих значень параметра рівня підготовленості нескінченна, і вони неперервно заповнюють всю позитивну піввісь. З цих простих міркувань випливає, що учасники тестування, що набрали однакову кількість первинних балів, майже напевно мають різні істинні рівні підготовленості. Але «відчути» ці відмінності ми не в змозі при тій точності оцінок, яка обумовлена даним об'ємом початкової інформації.

Зауваження 2. Аналогічні міркування можна провести і для первинних балів завдань. Оцінки рівнів складності збігаються для всіх завдань, що характеризуються однаковими первинними балами. Отже, можна упорядкувати всі завдання тесту в порядку зростання оцінок їх складності тільки на підставі первинних балів завдань. Але, очевидно, що тут ситуація більш благополучна. У будь-якому випадку, для підвищення нашої «чутливості» необхідно або якісно змінювати структуру матриці відповідей (тобто міняти модель і технологію тестування), або збільшувати її розмірність (зокрема, збільшувати число завдань k в тесті).

Покажемо, що у моделі Partial Credit первинні бали теж є достатніми статистиками. Елементами матриці відповідей тут будуть числа x_{ij}, які залежать від кількості категорій у завданнях. Ймовірність того, що i-ий учасник матиме конкретний рядок відповідей $x_{i1}, x_{i2}, \ldots, x_{ik}$ на основі теореми множення ймовірностей відповідно до (1.23) матиме вигляд:

$$P\left\{(x_{ij})\middle|\theta_i,(\delta_{jl})\right\}=\prod_{j=1}^{k}\left[\frac{\exp\sum_{g=0}^{x_{ij}}(\theta_i-\delta_{jg})}{\sum_{h=0}^{m_j}\exp\sum_{g=0}^{h}(\theta_i-\delta_{jg})}\right]=\frac{\exp\sum_{j=1}^{k}\sum_{g=0}^{x_{ij}}(\theta_i-\delta_{jg})}{\Phi_i},$$ де

$\Phi_i=\prod_{j=1}^{k}\left[\sum_{h=0}^{m_j}\exp\sum_{g=0}^{h}(\theta_i-\delta_{jg})\right]$. Ймовірність того, що i-та особа отримає рівно b_i первинних балів знаходимо, додаючи ймовірності всіх можливих реалізацій, які дають суму b_i:

$$P\left\{b_i \,\middle|\, \theta_i, (\delta_{jl})\right\} = \sum_{(b_i)} \frac{\exp\sum_{j=1}^{k}\sum_{g=0}^{x_{ij}}(\theta_i - \delta_{jg})}{\Phi_i} = \frac{\exp(b_i\theta_i)}{\Phi_i} \sum_{(b_i)} \exp\left(-\sum_{j=1}^{k}\sum_{g=0}^{x_{ij}}\delta_{jg}\right).$$

Остаточно запишемо умовну ймовірність того, що учасник матиме конкретну реалізацію $x_{i1}, x_{i2}, \ldots, x_{ik}$, якщо сума його балів b_i:

$$P\left\{(x_{ij}) \,\middle|\, b_i, (\delta_{jl})\right\} = \frac{P\left\{(x_{ij}) \,\middle|\, \theta_i, (\delta_{jl})\right\}}{P\left\{b_i \,\middle|\, \theta_i, (\delta_{jl})\right\}} = \frac{\exp\left(-\sum_{j=1}^{k}\sum_{g=0}^{x_{ij}}\delta_{jg}\right)}{\sum_{(b_i)} \exp\left(-\sum_{j=1}^{k}\sum_{g=0}^{x_{ij}}\delta_{jg}\right)}. \quad (2.27)$$

Ця умовна ймовірність, як і у моделі Раша, не залежить від рівня підготовки учасників тестування, а лише від складностей кроків усіх завдань. Отже, первинні бали є достатніми статистиками для параметрів підготовленості, а тому при оцінюванні параметрів можна скоротити об'єм початкових даних. Така редукція початкових даних теоретично повністю обгрунтована лише в рамках певних моделей, зокрема, в рамках моделей сімейства Раша. В загальному випадку така редукція може виявитися некоректною.

Аналогічно доводиться така властивість відокремлення для первинних балів завдань.

2.3. Метод моментів оцінки латентних параметрів

Задачею сучасної теорії тестів є не лише адекватне моделювання процесу тестування, але і розробка методів оцінювання латентних параметрів за наявним статистичним матеріалом, отриманим у результаті тестування. Якщо існує деякий алгоритм обчислення значень латентних параметрів за відомою матрицею відповідей $A = (a_{ij})$, то зрозуміло, що отримані оцінки $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_n$ та $\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_k$ відрізняються від їх точних значень θ_i та δ_j. Критерієм якості оцінок є відомі зі статистики їх характеристики: незміщеність, ефективність та слушність. На сьогодні розроблено багато різних методів побудови оцінок латентних параметрів, але не всі вони дозволяють отримати оцінки, що володіють усіма необхідними властивостями.

Найпростіший спосіб, який дає слушні, але не ефективні оцінки, ґрунтується на формулі (1.14) та на використанні частот як оцінок відповідних ймовірностей. Нагадаємо, що у моделі Раша

$$l_{ij} = \ln\left(\frac{p_{ij}}{q_{ij}}\right) = \theta_i - \delta_j, \qquad (2.28)$$

де $p_{ij} \neq 0$ та $p_{ij} \neq 1$ (асимптоти логістичної кривої).

Позначимо n_b – кількість учасників тестування, які набрали один і той же первинний бал $b = 0,1,2,\ldots,k$, $\sum_{b=0}^{k} n_b = n$; $\theta_i(b)$ – рівень підготовки *i*-го учасника, що має бал b. Як було доведено вище, всі учасники з однаковим балом у моделі Раша повинні мати однакові оцінки рівня підготовленості $\hat{\theta}(b)$. Нехай $\hat{p}_j(b)$ – відносна частота правильної відповіді на *j*–те запитання серед учасників, які мають однаковий бал b. Вона легко обчислюється за відомою матрицею відповідей і є незміщеною та слушною оцінкою відповідної ймовірності. Тоді $\hat{q}_j(b) = 1 - \hat{p}_j(b)$, $\hat{l}_j(b) = \ln\left(\frac{\hat{p}_j(b)}{\hat{q}_j(b)}\right)$.

Перепишемо рівність (2.28) для оцінок:

$$\hat{\theta}(b) - \hat{\delta}_j = \hat{l}_j(b); \quad j = 1,2,\ldots,k; \quad b = 0,1,2,\ldots,k. \qquad (2.29)$$

Праві частини цих рівностей вважаємо відомими. Щоб уникнути обчислення логарифмів *0* та ∞, потрібно відкоректувати екстремальні значення $\hat{p}_j(b) = 1$ та $\hat{p}_j(b) = 0$ (коли серед учасників з балом b всі відповіли вірно (або не вірно) на *j*-те запитання). Для цього достатньо зменшити (або збільшити) екстремальне значення на деяку малу сталу Δ. Якщо, наприклад, $\Delta = \frac{1}{150}$, то $\ln\frac{1-\Delta}{\Delta} \approx 5$, а $\ln\frac{\Delta}{1-\Delta} \approx -5$. А більша різниця між параметрами малоймовірна (зауваження 1 пункту 1.2).

Формули (2.29) визначають систему $k \cdot (k+1)$ рівнянь з $2k+1$ невідомими $\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_k$ та $\hat{\theta}(0), \hat{\theta}(1), \ldots, \hat{\theta}(b)$:

$\hat{\delta}_1$	$\hat{\delta}_2$	…	$\hat{\delta}_k$	$\hat{\theta}(0)$	$\hat{\theta}(1)$	$\hat{\theta}(2)$	…	$\hat{\theta}(k)$	
-1				1					$\hat{l}_1(0)$
-1					1				$\hat{l}_1(1)$
-1						1			$\hat{l}_1(2)$
…							…		…
-1								1	$\hat{l}_1(k)$
	-1			1					$\hat{l}_2(0)$
	-1				1				$\hat{l}_2(1)$
	-1					1			$\hat{l}_2(2)$
	…						…		…
	-1							1	$\hat{l}_2(k)$
									…
			-1	1					$\hat{l}_k(0)$
			-1		1				$\hat{l}_k(1)$
			-1			1			$\hat{l}_k(2)$
			…				…		…
			-1					1	$\hat{l}_k(k)$

У загальному випадку кількість рівнянь перевищує кількість невідомих. Але матриця коефіцієнтів дуже розріджена, можна довести, що вона вироджена і її ранг дорівнює $2k$. Тому одну із невідомих змінних можна вибрати довільно (вона визначатиме початок на шкалі логітів), інші через неї виразити. Найчастіше початок шкали суміщають із середнім значенням $\overline{\theta}$ параметра θ в логітах. Методи розв'язування таких систем тут не розглядаються.

Покращення оцінок вимагає ускладнення алгоритму їх побудови. Одним із методів побудови точкових оцінок параметрів заданого розподілу є метод моментів Пірсона, який полягає у прирі-

внюванні теоретичних моментів даного розподілу до відповідних емпіричних моментів такого ж порядку. Прирівнювати можна не лише самі моменти, а й функції від них.

Прирівняємо первинні бали (емпіричні моменти першого порядку) до відповідних математичних сподівань (теоретичних моментів першого порядку), які для дихотомічних завдань обчислюються за формулами (2.5):

$$\begin{cases} b_i = M\{b_i\} = \sum_{j=1}^{k} p_{ij}, \quad i = 1,2,\ldots,n, \\ c_j = M\{c_j\} = \sum_{i=1}^{n} p_{ij}, \quad j = 1,2,\ldots,k. \end{cases}$$

Або

$$\begin{cases} b_i - \sum_{j=1}^{k} p_{ij} = 0, \quad i = 1,2,\ldots,n, \\ c_j - \sum_{i=1}^{n} p_{ij} = 0, \quad j = 1,2,\ldots,k. \end{cases} \tag{2.30}$$

Невідомі θ_i та δ_j входять до виразу p_{ij} (залежно від моделі). Система (2.30) у загальному випадку містить $n+k$ рівнянь з $n+k$ невідомими.

Якщо врахувати, що у моделі Раша всі учасники з однаковим балом повинні мати однакові оцінки рівня підготовленості, то з перших *n* рівнянь залишиться лише $k+1$ рівняння для балів від *0* до *k* (вилучаємо однакові рівняння). Якщо θ_b рівень підготовленості учасника із групи, яка набрала b балів ($b = 0,1,2,\ldots,k$), то ймовірність того, що такий учасник вірно виконає *j*–те завдання відповідно до моделі Раша (1.10) має вигляд

$$p_{bj} = \frac{e^{\theta_b - \delta_j}}{1 + e^{\theta_b - \delta_j}}. \tag{2.31}$$

У цьому випадку система (2.30) матиме такий розгорнутий вигляд (n_b – кількість учасників з балом b):

$$
\begin{cases}
\dfrac{e^{\theta_0-\delta_1}}{1+e^{\theta_0-\delta_1}}+\dfrac{e^{\theta_0-\delta_2}}{1+e^{\theta_0-\delta_2}}+\dfrac{e^{\theta_0-\delta_3}}{1+e^{\theta_0-\delta_3}}+\ldots\dfrac{e^{\theta_0-\delta_k}}{1+e^{\theta_0-\delta_k}}=0, \\
\dfrac{e^{\theta_1-\delta_1}}{1+e^{\theta_1-\delta_1}}+\dfrac{e^{\theta_1-\delta_2}}{1+e^{\theta_1-\delta_2}}+\dfrac{e^{\theta_1-\delta_3}}{1+e^{\theta_1-\delta_3}}+\ldots\dfrac{e^{\theta_1-\delta_k}}{1+e^{\theta_1-\delta_k}}=1, \\
\ldots\ldots\ldots\ldots\ldots \\
\dfrac{e^{\theta_k-\delta_1}}{1+e^{\theta_k-\delta_1}}+\dfrac{e^{\theta_k-\delta_2}}{1+e^{\theta_k-\delta_2}}+\dfrac{e^{\theta_k-\delta_3}}{1+e^{\theta_k-\delta_3}}+\ldots\dfrac{e^{\theta_k-\delta_k}}{1+e^{\theta_k-\delta_k}}=k, \\
n_0\cdot\dfrac{e^{\theta_0-\delta_1}}{1+e^{\theta_0-\delta_1}}+n_1\cdot\dfrac{e^{\theta_1-\delta_1}}{1+e^{\theta_1-\delta_1}}+n_2\cdot\dfrac{e^{\theta_2-\delta_1}}{1+e^{\theta_2-\delta_1}}+\ldots n_k\cdot\dfrac{e^{\theta_k-\delta_1}}{1+e^{\theta_k-\delta_1}}=c_1, \\
n_0\cdot\dfrac{e^{\theta_0-\delta_2}}{1+e^{\theta_0-\delta_2}}+n_1\cdot\dfrac{e^{\theta_1-\delta_2}}{1+e^{\theta_1-\delta_2}}+n_2\cdot\dfrac{e^{\theta_2-\delta_2}}{1+e^{\theta_2-\delta_2}}+\ldots n_k\cdot\dfrac{e^{\theta_k-\delta_2}}{1+e^{\theta_k-\delta_2}}=c_2, \\
\ldots\ldots\ldots\ldots\ldots \\
n_0\cdot\dfrac{e^{\theta_0-\delta_k}}{1+e^{\theta_0-\delta_k}}+n_1\cdot\dfrac{e^{\theta_1-\delta_k}}{1+e^{\theta_1-\delta_k}}+n_2\cdot\dfrac{e^{\theta_2-\delta_k}}{1+e^{\theta_2-\delta_k}}+\ldots n_k\cdot\dfrac{e^{\theta_k-\delta_k}}{1+e^{\theta_k-\delta_k}}=c_k.
\end{cases}
$$

Система має структуру, яка дозволяє послідовно знаходити розв'язки для θ_i, якби були відомі складності завдань δ_j, і навпаки. Причому, для цього необхідно розв'язати лише одне рівняння. Наприклад, якщо уявити, що відомі складності усіх завдань δ_j, то перше нелінійне трансцендентне рівняння, яке містить лише одну невідому θ_0, можна розв'язати відносно θ_0, використовуючи один з наближених методів, наприклад, метод Ньютона. Аналогічно з кожного наступного рівняння послідовно можна знайти $\theta_1,\theta_2,\ldots,\theta_k$. Початкові наближення рівнів підготовки $\theta_b^{(0)}$ можна оцінити на основі матриці відповідей як логарифм відношення шансів на успіх $\theta_b^{(0)}=\ln\dfrac{b}{k-b}$. Навпаки, якщо відомі рівні підготовки всіх учасників θ_i, то з першого рівняння другої групи знаходимо δ_1 (для цього потрібне початкове значення $\delta_1^{(0)}$), далі $\delta_2,\delta_3,\ldots,\delta_k$ (початкові значення $\delta_2^{(0)},\delta_3^{(0)},\ldots,\delta_k^{(0)}$ оцінюються теж за матрицею відповідей).

Нагадаємо розрахункову ітераційну формулу методу Ньютона (дотичних) для знаходження розв'язку нелінійного рівняння

виду $f(x)=0$, де $f(x)$ – двічі диференційовна функція, а $x^{(0)}$ – початкове наближення шуканого кореня:

$$x^{(\nu+1)} = x^{(\nu)} - \frac{f(x^{(\nu)})}{f'(x^{(\nu)})}, \tag{2.32}$$

де ν – номер послідовного наближення.

Похідні лівих частин усіх рівнянь системи (2.30) з урахуванням проведеної редукції мають вигляд (тут $q_{bj}=1-p_{bj}$):

$$\frac{\partial}{\partial\theta_b}\left(b-\sum_{j=1}^{k}p_{bj}\right)=-\sum_{j=1}^{k}p_{bj}\cdot q_{bj},$$

$$\frac{\partial}{\partial\delta_j}\left(-c_j+\sum_{b=0}^{k}n_b p_{bj}\right)=-\sum_{b=0}^{k}n_b\cdot p_{bj}\cdot q_{bj}.$$

Ітераційні формули (2.32) для кожного з рівнянь системи (2.30) можуть бути записані у вигляді:

$$\theta_b^{(\nu+1)} = \theta_b^{(\nu)} + \frac{b-\sum_{j=1}^{k}p_{bj}^{(\nu)}}{\sum_{j=1}^{k}p_{bj}^{(\nu)}\cdot q_{bj}^{(\nu)}}, \quad b=0,1,2,\ldots,k, \tag{2.33}$$

$$\delta_j^{(\mu+1)} = \delta_j^{(\mu)} + \frac{-c_j+\sum_{b=0}^{k}n_b\cdot p_{bj}^{(\mu)}}{\sum_{b=0}^{k}n_b\cdot p_{bj}^{(\mu)}\cdot q_{bj}^{(\mu)}}, \quad j=1,2,\ldots,k. \tag{2.34}$$

Тут b – номер групи учасників тестування, які набрали один і той же первинний бал b, n_b – кількість учасників у цій групі, $p_{bj}^{(\nu)}$ та $p_{bj}^{(\mu)}$ – ймовірності, що обчислюються за формулами (2.31) за наближеними значеннями латентних параметрів на даному кроці. При цьому ймовірності, що відповідають екстремальним балам $b=0$ та $b=k$ потрібно відкоректувати, як це робили у попередньому алгоритмі. Тобто, $p_{0j}=\Delta$, $p_{kj}=1-\Delta$, де Δ – деяка мала додатна стала, наприклад $\Delta=1/150$.

Отже, основні кроки алгоритму розв'язування системи (2.30)

для знаходження оцінок латентних параметрів такі:

1. Покладемо $\nu = 0$ і обчислимо початкові наближення $\theta_b^{(0)}$ для кожного b: $\theta_b^{(0)} = \ln\dfrac{b}{k-b}$ для всіх $b = 1, 2, \ldots, k-1$;

$$\theta_0^{(0)} = \ln\frac{\Delta}{1-\Delta}, \quad \theta_k^{(0)} = \ln\frac{1-\Delta}{\Delta}.$$

Знаходимо середнє значення $\bar{\theta}^{(0)} = \dfrac{1}{n}\sum_{b=0}^{k} n_b \cdot \theta_b^{(0)}$ і центруємо оцінки $\theta_b^{(0)}$, тобто обчислюємо $\dot{\theta}_b^{(0)} = \theta_b^{(0)} - \bar{\theta}^{(0)}$.

2. Покладемо $\mu = 0$ і обчислимо початкові наближення $\delta_j^{(0)}$ для кожного j: $\delta_j^{(0)} = \ln\dfrac{n-c_j}{n}$, $j = 1, 2, \ldots, k$.

3. Обчислюємо чергові наближення $\delta_j^{(\mu+1)}$ за формулою (2.34), де $p_{bj}^{(\mu)} = \left(1 + \exp\left(-(\dot{\theta}_b - \delta_j^{(\mu)})\right)\right)^{-1}$, поки не буде виконуватись нерівність $\left|\delta_j^{(\mu+1)} - \delta_j^{(\mu)}\right| < \varepsilon$ для всіх $j = 1, 2, \ldots, k$.

Тут $\dot{\theta}_b$ – центрована оцінка рівня підготовленості, яка була обчислена до цього кроку, ε – деяка мала додатна стала, що визначає точність вимірювання.

4. Обчислюємо чергові наближення $\theta_b^{(\nu+1)}$ за формулою (2.33), де $p_{bj}^{(\nu)} = \left(1 + \exp\left(-(\dot{\theta}_b^{(\nu)} - \delta_j)\right)\right)^{-1}$, поки не буде виконуватись нерівність $\left|\theta_b^{(\nu+1)} - \theta_b^{(\nu)}\right| < \varepsilon$ для всіх $b = 1, 2, \ldots, k-1$.

Тут δ_j – оцінка складності завдань, яка була обчислена перед цим кроком.

5. Знаходимо середнє значення $\bar{\theta}^{(\nu)} = \dfrac{1}{n - n_0 - n_k}\sum_{b=1}^{k-1} n_b \cdot \theta_b^{(\nu)}$ і центруємо оцінки $\theta_b^{(\nu)}$, тобто обчислюємо $\dot{\theta}_b^{(\nu)} = \theta_b^{(\nu)} - \bar{\theta}^{(\nu)}$.

6. Обчислюємо середнє квадратичне відхилення оцінок чергового наближення від аналогічних оцінок попереднього наближення

$$\sigma = \frac{1}{\sqrt{2k-1}}\left[\sum_{b=1}^{k-1}\left(\dot{\theta}_b^{(\nu+1)} - \dot{\theta}^{(\nu)}\right)^2 + \sum_{j=1}^{k}\left(\delta_j^{(\mu+1)} - \delta_j^{(\mu)}\right)^2\right]^{\frac{1}{2}}$$

Якщо $\sigma > \frac{\varepsilon}{3}$, то переходимо до пункту 3. У іншому випадку обчислення закінчуємо, оскільки досягнута бажана точність.

Отримані оцінки є слушними, асимптотично незміщеними, ефективними та асимптотично ефективними.

2.4. Метод максимальної вірогідності

Для оцінки латентних параметрів θ_i та δ_j за відомою матрицею відповідей можна використати метод максимальної вірогідності, запропонований у 1912 році англійським статистиком Р.Фішером. Відомо, що оцінки, отримані за методом максимальної вірогідності, є слушними, асимптотично незміщеними та асимптотично ефективними.

Функція вірогідності L залежить від невідомих параметрів θ_i та δ_j і , в силу припущення про локальну незалежність, дорівнює добутку ймовірностей для всіх можливих значень *i* та *j*. Для дихотомічних завдань, коли $p_{ij} = p(\theta_i, \delta_j)$ – ймовірність для *i*-го учасника відповісти вірно на *j*-те запитання, a_{ij} – елементи матриці, що складається з 0 та 1, функція вірогідності має вигляд:

$$L(\theta_i, \delta_j) = \prod_{i=1}^{n}\prod_{j=1}^{k} p\{a_{ij}\,|\,(\theta_i, \delta_j)\} = \prod_{i=1}^{n}\prod_{j=1}^{k} p_{ij}^{\,a_{ij}}\left(1 - p_{ij}\right)^{1-a_{ij}} .$$

Значення $\hat{\theta}_i$ та $\hat{\delta}_j$, при яких функція L досягає максимуму, будуть точковими оцінками максимальної вірогідності невідомих параметрів θ_i та δ_j. Далі зручно досліджувати функцію $\ln L$, логарифмічну функцію вірогідності, оскільки вона досягає максимуму при тих же значеннях аргументів, що і L:

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{k}\left[a_{ij}\ln p_{ij} + (1 - a_{ij})\ln(1 - p_{ij})\right]. \qquad (2.35)$$

Необхідною умовою екстремуму є рівність нулеві усіх частинних похідних. Знайдемо та прирівняємо до нуля частинні похі-

дні по θ_i та δ_j функції $\ln L$:

$$\begin{cases} \dfrac{\partial(\ln L)}{\partial\theta_i} = \sum\limits_{j=1}^{k} \dfrac{a_{ij} - p_{ij}}{p_{ij}(1-p_{ij})} \cdot \dfrac{\partial p_{ij}}{\partial\theta_i} = 0, \quad i = 1,2,\ldots,n; \\ \dfrac{\partial(\ln L)}{\partial\delta_j} = \sum\limits_{i=1}^{n} \dfrac{a_{ij} - p_{ij}}{p_{ij}(1-p_{ij})} \cdot \dfrac{\partial p_{ij}}{\partial\delta_j} = 0, \quad j = 1,2,\ldots,k. \end{cases} \tag{2.36}$$

Отримана система (2.36) називається *системою рівнянь вірогідності*. Для моделі Раша (або 1PL) це система $n+k$ рівнянь з $n+k$ невідомими. Для інших моделей Бірнбаума 2PL та 3PL потрібно додати ще частинні похідні по параметрах диференціюючої здатності та угадування, кожна додає ще по k рівнянь. Розмірність даної системи можна зменшити, якщо врахувати, як і раніше, що різні оцінки параметра θ_i можуть відповідати лише різним первинним балам.

Для моделі Раша

$$\frac{\partial p_{ij}}{\partial\theta_i} = \frac{\partial}{\partial\theta_i}\left(\frac{e^{\theta_i-\delta_j}}{1+e^{\theta_i-\delta_j}}\right) = \frac{e^{\theta_i-\delta_j}}{\left(1+e^{\theta_i-\delta_j}\right)^2} = p_{ij}\cdot\left(1-p_{ij}\right),$$

$$\frac{\partial p_{ij}}{\partial\delta_j} = \frac{\partial}{\partial\delta_j}\left(\frac{e^{\theta_i-\delta_j}}{1+e^{\theta_i-\delta_j}}\right) = -\frac{e^{\theta_i-\delta_j}}{\left(1+e^{\theta_i-\delta_j}\right)^2} = -p_{ij}\cdot\left(1-p_{ij}\right).$$

Тому система (2.36) матиме вигляд:

$$\begin{cases} \sum\limits_{j=1}^{k}\left(a_{ij} - p_{ij}\right) = b_i - \sum\limits_{j=1}^{k} p_{ij} = 0, \quad i = 1,2,\ldots n; \\ -\sum\limits_{i=1}^{n}\left(a_{ij} - p_{ij}\right) = -c_j + \sum\limits_{i=1}^{n} p_{ij} = 0, \quad j = 1,2,\ldots k. \end{cases}$$

Ця система повністю збігається із системою (2.30) попереднього пункту. Далі для її розв'язання використовуємо розглянутий вище ітераційний метод. Для інших дихотомічних моделей система вірогідності відрізнятиметься від системи (2.30), яку пропонує метод моментів для всіх моделей.

Побудуємо систему вірогідності для моделі з політомічними завданнями Partial Credit. Нехай x_{ij} – кількість балів, які отримує i-та особа у j–му завданні, $p_{x_{ij}ij}$ – відповідна ймовірність, яка об-

числюється за формулою (1.23). Тоді функція вірогідності L має вигляд:

$$L(\theta_i, \delta_{jl}) = \prod_{i=1}^{n}\prod_{j=1}^{k} p_{x_{ij}ij} = \frac{\exp\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{g=0}^{x_{ij}}\left(\theta_i - \delta_{jg}\right)}{\prod_{i=1}^{n}\prod_{j=1}^{k}\left[\sum_{h=0}^{m_j}\exp\sum_{g=0}^{h}\left(\theta_i - \delta_{jg}\right)\right]}.$$

На екстремум дослідимо логарифм цієї функції:

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}\theta_i - \sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{g=1}^{x_{ij}}\delta_{jg} - \sum_{i=1}^{n}\sum_{j=1}^{k}\ln\left[\sum_{h=0}^{m_j}\exp\sum_{g=0}^{h}\left(\theta_i - \delta_{jg}\right)\right]. \qquad (2.37)$$

Тут $\sum_{g=0}^{x_{ij}}\delta_{jg} = \sum_{g=1}^{x_{ij}}\delta_{jg}$, оскільки $\delta_{j0} \equiv 0$. Функцію (2.37) можна дещо спростити, якщо врахувати, що первинний бал *i*-го учасника $b_i = \sum_{j=1}^{k} x_{ij}$. Також зауважимо, що $\sum_{g=1}^{x_{ij}}\delta_{jg}$ – сума складностей кроків, завершених *i*-ою особою у *j*–му завданні, тоді $\sum_{i=1}^{n}\sum_{g=1}^{x_{ij}}\delta_{jg}$ – сума складностей завершених кроків у *j*–му завданні всіма учасниками тестування. Позначимо s_{jg} – кількість учасників, які завершать крок g у *j*–му завданні (таблиця 2). Тоді попередню суму можна переписати $\sum_{i=1}^{n}\sum_{g=1}^{x_{ij}}\delta_{jg} = \sum_{g=1}^{m_j} s_{jg}\delta_{jg}$. Після спрощень маємо:

$$\ln L = \sum_{i=1}^{n} b_i\theta_i - \sum_{j=1}^{k}\sum_{g=1}^{m_j} s_{jg}\delta_{jg} - \sum_{i=1}^{n}\sum_{j=1}^{k}\ln\left[\sum_{h=0}^{m_j}\exp\sum_{g=0}^{h}\left(\theta_i - \delta_{jg}\right)\right]. \qquad (2.38)$$

Знайдемо похідні від логарифма квадратних дужок у (2.38):

$$\frac{\partial}{\partial\theta_i}\left(\ln\left[\sum_{h=0}^{m_j}\exp\sum_{g=0}^{h}\left(\theta_i - \delta_{jg}\right)\right]\right) = \frac{\sum_{h=0}^{m_j} h\cdot\exp\sum_{g=0}^{h}\left(\theta_i - \delta_{jg}\right)}{\sum_{h=0}^{m_j}\exp\sum_{g=0}^{h}\left(\theta_i - \delta_{jg}\right)} =$$

$$= \sum_{h=0}^{m_j} h \cdot p_{hij} = \sum_{h=1}^{m_j} h \cdot p_{hij} \; ;$$

$$\frac{\partial}{\partial \delta_{jg}} \left(\ln \left[\sum_{h=0}^{m_j} \exp \sum_{r=0}^{h} \left(\theta_i - \delta_{jr} \right) \right] \right) = \frac{- \sum_{h=g}^{m_j} \exp \sum_{r=0}^{h} \left(\theta_i - \delta_{jr} \right)}{\sum_{h=0}^{m_j} \exp \sum_{r=0}^{h} \left(\theta_i - \delta_{jr} \right)} =$$

$$= - \sum_{h=g}^{m_j} p_{hij} \; .$$

Оскільки δ_{jg} з'являється лише у доданках, для яких $h \geq g$, то похідна містить доданки, починаючи від *g*. Тепер прирівняємо до нуля частинні похідні функції (2.38):

$$\begin{cases} \dfrac{\partial (\ln L)}{\partial \theta_i} = b_i - \sum\limits_{j=1}^{k} \sum\limits_{h=1}^{m_j} h \cdot p_{hij} = 0, \quad i = 1, 2, \ldots, n; \\ \dfrac{\partial (\ln L)}{\partial \delta_{jg}} = -s_{jg} + \sum\limits_{i=1}^{n} \sum\limits_{h=g}^{m_j} p_{hij} = 0, \quad j = 1, 2, \ldots, k; \; g = 1, 2, \ldots, m_j . \end{cases} \tag{2.39}$$

Тут $\sum\limits_{h=1}^{m_j} h \cdot p_{hij}$ – очікуваний бал (математичне сподівання) *i-г*о учасника у *j*–му завданні, а $\sum\limits_{j=1}^{k} \sum\limits_{h=1}^{m_j} h \cdot p_{hij}$ – очікуваний бал за весь тест. Аналогічно, у другому рівняння фігурує очікувана кількість осіб, що завершать не менше g кроків у *j*–му завданні. Тобто, методом моментів отримали б таку ж систему.

Якщо врахувати, що особи з рівними первинними балами мають однакові оцінки рівня підготовленості, то у системі (2.39) замість перших *n* рівнянь залишиться (*M*+1) рівняння, де *M* – максимально можливий первинний бал $M = \sum\limits_{j=1}^{k} m_j$, і нова система (2*M*+1) рівнянь з (2*M*+1) невідомими $\theta_0, \ldots, \theta_M, \delta_{11}, \ldots, \delta_{km_k}$ матиме вигляд:

$$\begin{cases} b - \sum_{j=1}^{k} \sum_{h=1}^{m_j} h \cdot p_{hbj} = 0, \quad b = 0,1,2,\ldots,M; \\ s_{jg} - \sum_{b=0}^{M} n_b \sum_{h=g}^{m_j} p_{hbj} = 0, \quad j = 1,2,\ldots,k;\ g = 1,2,\ldots,m_j. \end{cases} \tag{2.40}$$

Дана система має єдиний розв'язок, до неї аналогічно застосовуємо ітераційний метод Ньютона з послідовними ітераціями:

$$\theta_b^{(\nu+1)} = \theta_b^{(\nu)} + \frac{b - \sum_{j=1}^{k} \sum_{h=1}^{m_j} h \cdot p_{hbj}^{(\nu)}}{\sum_{j=1}^{k} \left[\sum_{h=1}^{m_j} h^2 p_{hbj}^{(\nu)} - \left(\sum_{h=1}^{m_j} h \cdot p_{hbj}^{(\nu)} \right)^2 \right]}, \quad b = \overline{0, M}, \tag{2.41}$$

$$\delta_{jg}^{(\mu+1)} = \delta_{jg}^{(\mu)} + \frac{-s_{jg} + \sum_{b=0}^{M} n_b \sum_{h=g}^{m_j} p_{hbj}^{(\mu)}}{\sum_{b=0}^{M} n_b \cdot \left[\sum_{h=g}^{m_j} p_{hbj}^{(\mu)} - \left(\sum_{h=g}^{m_j} p_{hbj}^{(\mu)} \right)^2 \right]}, \ j = \overline{1,k};\ g = \overline{1,m_j}. \tag{2.42}$$

Тут $p_{hbj}^{(\nu)}$ та $p_{hbj}^{(\mu)}$ – ймовірності, які обчислюються за формулою (1.23) за відомими на той момент значеннями латентних параметрів. Початкові наближення вибираємо з умов $\theta_b^{(0)} = \ln \frac{b}{M-b}$ та $\delta_{jg}^{(0)} = \ln \frac{s_{j(g-1)}}{s_{jg}}$. Далі будуємо розв'язок за алгоритмом 1–6. Ітераційний процес завершується, коли виконується нерівність

$$\left[\sum_{b=0}^{M} \left(\dot{\theta}_b^{(\nu+1)} - \dot{\theta}^{(\nu)} \right)^2 + \sum_{j=1}^{k} \sum_{g=1}^{m_j} \left(\delta_{jg}^{(\mu+1)} - \delta_{jg}^{(\mu)} \right)^2 \right]^{\frac{1}{2}} < \varepsilon .$$

Зауваження 1. Оцінки, отримані методом максимальної вірогідності, є слушними, асимптотично незміщеними та асимптотично ефективними. Вони мають асимптотично нормальний розподіл з параметрами:

$$M(\hat{\theta}) = \theta,\ M(\hat{\delta}) = \delta,\ D(\hat{\theta}) \approx \left[\frac{\partial^2 (\ln L)}{\partial \theta^2} \right]^{-1}_{\hat{\theta},\hat{\delta}},\ D(\hat{\delta}) \approx \left[\frac{\partial^2 (\ln L)}{\partial \delta^2} \right]^{-1}_{\hat{\theta},\hat{\delta}}.$$

Причому ця дисперсія мінімальна у порівнянні з іншими асимптотично нормальними оцінками. Тобто, якщо існують ефективні оцінки, то метод максимальної вірогідності їх дає. Враховуючи вигляд другої похідної логарифма функції вірогідності (це якраз знаменники ітераційних формул (2.33), (2.34), (2.41) та (2.42)), можемо записати формули для *стандартних похибок* вимірювання латентних параметрів у моделі Раша

$$SE(\hat{\theta}_i) = \left[\sum_{j=1}^{k} p_{ij} q_{ij} \right]^{-\frac{1}{2}},\ \ SE(\hat{\delta}_j) = \left[\sum_{i=1}^{n} p_{ij} q_{ij} \right]^{-\frac{1}{2}}.$$

Для моделі Partial Credit маємо:

$$SE(\hat{\theta}_i) = \left(\sum_{j=1}^{k} \left[\sum_{h=1}^{m_j} h^2 p_{hij} - \left(\sum_{h=1}^{m_j} h \cdot p_{hij} \right)^2 \right] \right)^{-\frac{1}{2}},$$

$$SE(\hat{\delta}_{jg}) = \left(\sum_{i=1}^{n} \left[\sum_{h=g}^{m_j} p_{hij} - \left(\sum_{h=g}^{m_j} p_{hij} \right)^2 \right] \right)^{-\frac{1}{2}}.$$

Отже, метод максимальної вірогідності має сенс винятково для великих вибірок випробуваних (не менш 200-300 випробуваних) і досить довгих тестів (не менше 30 завдань). В цілому ж з точки зору теорії такі оцінки латентних параметрів є найбільш ефективними і можуть бути прийняті за їх дійсні значення. Звичайно, для реалізації методу максимальної вірогідності потрібні спеціальні програми. Якщо використовується банк тестових завдань з відомими стійкими характеристиками складності, то можна отримати оцінки рівня підготовленості з мінімальною стандартною помилкою вимірювання.

Зауваження 2. Припустимо, що всі параметри завдань уже відомі. Тоді для знаходження оцінок рівнів підготовленості для будь-якої моделі можна використовувати перше рівняння системи (2.36), яке перепишемо у такому вигляді:

$$\sum_{j=1}^{k} \omega_j(\theta)\, a_{ij} = \sum_{j=1}^{k} \omega_j(\theta)\, P_j(\theta), \qquad (2.43)$$

де $P_j(\theta)$ – ймовірність правильної відповіді учасника з рівнем підготовленості θ на *j*–те завдання тесту. Множники

$$\omega_j(\theta) = \frac{P_j'(\theta)}{P_j(\theta) \cdot \left(1 - P_j(\theta)\right)} \tag{2.44}$$

називаються *оптимальними ваговими коефіцієнтами* завдання *j*. Тоді формула (2.43) виражає рівність зваженого первинного балу та зваженого очікуваного балу.

Для моделі Раша вагові коефіцієнти завдань $\omega_j(\theta) = 1$ (обчислені за формулою (2.44)) і не залежать від θ. У цій моделі нема потреби присвоювати завданням ваги. Для двопараметричної моделі 2PL обчислення дають ваги $\omega_j(\theta) = Dd_j$, які теж не залежать від рівня підготовленості. У цій моделі внесок кожного завдання у підсумковий бал повинен бути пропорційним роздільній здатності завдання. Перш ніж будувати оцінки для θ, у формулі (2.43) потрібно первинні бали підсилити вагами завдань $D\sum_{j=1}^{k} d_j \cdot a_{ij}$. Для трипараметричної моделі вагові коефіцієнти залежать від θ:

$$\omega_j(\theta) = \frac{Dd_j}{\left(1 - c_j\right)} \cdot \frac{P_j(\theta) - c_j}{P_j(\theta)}, \tag{2.45}$$

де $P_j(\theta)$ – ймовірність правильної відповіді у моделі 3PL, що обчислюється за формулою (1.17). Можна показати, що для сильних учасників вагові коефіцієнти прямують до d_j, оскільки ймовірність правильної відповіді на питання для таких опитаних прямує до 1: $\lim_{\theta \to \infty} \omega_j(\theta) = \lim_{P_j(\theta) \to 1} \frac{Dd_j}{\left(1 - c_j\right)} \cdot \frac{P_j(\theta) - c_j}{P_j(\theta)} = \frac{Dd_j}{\left(1 - c_j\right)} \cdot \frac{1 - c_j}{1} = Dd_j$. Для слабких учасників, навпаки, $P_j(\theta) \xrightarrow[\theta \to 0]{} c_j$, тому вагові коефіцієнти прямують до нуля: $\lim_{\theta \to 0} \omega_j(\theta) = \frac{Dd_j}{\left(1 - c_j\right)} \cdot \frac{c_j - c_j}{c_j} = 0$.

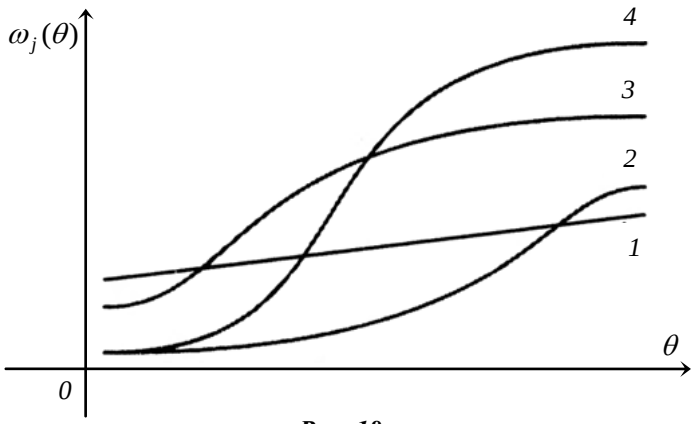


Рис. 18.

На рис.18 зображено вагові коефіцієнти чотирьох завдань. Крива 1 відповідає завданню з низькою диференційною здатністю ($d_1 < 0{,}5$), тому внесок такого завдання у загальну оцінку незначний як для сильних, так і для слабких опитаних. Криві 2 та 4 відповідають досить складним завданням (але $d_4 > d_2$), тому при малих θ внесок обох завдань у загальний бал слабких учасників близький до нуля. При великих θ вирішальний внесок робить параметр d_4. Завдання 3 з $d_3 \approx 1$ має не велику складність, тому вагові коефіцієнти для нього відмінні від нуля навіть для слабких учасників. Використання оптимальних вагових коефіцієнтів не набуло поширення, оскільки неможливе при сумісному оцінюванні латентних параметрів.

2.5. Метод умовної максимальної вірогідності

Процедуру методу максимальної вірогідності, яку розглянули вище, ще називають *Joint Maximum Likelihood procedure* (JML), оскільки вона дозволяє одночасно оцінювати латентні параметри учасників тестування та завдань. Розроблений також вдосконалений метод максимальної вірогідності *Marginal Maximum Likelihood procedure* (MML), який дозволяє відокремити оцінки учасників тестування від оцінок завдань. JML та MML процедури можуть використовуватись для всіх математичних моделей тестів.

Крім того, для однопараметричних моделей (типу Раша, Partial Credit) розроблений варіант методу максимальної вірогідності, який дозволяє проводити оцінку складності завдань окремо від оцінки рівнів підготовленості. Його називають *Conditional Maximum Likelihood procedure* (CML). Розглянемо ідею методу CML на прикладі моделі Раша.

Перепишемо для моделі Раша ймовірність правильної відповіді особи з рівнем підготовки θ на *j*-те питання у вигляді:

$$P_j(\theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)} = \frac{\varsigma\varepsilon_j}{1 + \varsigma\varepsilon_j}, \qquad (2.46)$$

де $\varsigma = \exp(\theta)$ – параметр учасника, $\varepsilon_j = \exp(\delta_j)$ – завдання. Ймовірність неправильної відповіді $1 - P_j(\theta) = 1 - \frac{\varsigma\varepsilon_j}{1 + \varsigma\varepsilon_j} = \frac{1}{1 + \varsigma\varepsilon_j}$.

Нехай є два завдання з параметрами ε_1 та ε_2. Умовна ймовірність того, що буде дана правильна відповідь на перше питання, якщо набраний бал $b = 1$ (тобто з двох можливих у даній ситуації векторів відповідей (1; 0) та (0; 1) реалізується перший) має вигляд:

$$P\{(1;0) \mid b = 1\} = \frac{\frac{\varsigma\varepsilon_1}{1 + \varsigma\varepsilon_1} \times \frac{1}{1 + \varsigma\varepsilon_2}}{\frac{\varsigma\varepsilon_1}{1 + \varsigma\varepsilon_1} \times \frac{1}{1 + \varsigma\varepsilon_2} + \frac{1}{1 + \varsigma\varepsilon_1} \times \frac{\varsigma\varepsilon_2}{1 + \varsigma\varepsilon_2}} = \frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2}.$$

Ця ймовірність не залежить від ς, тому ми можемо порівнювати завдання незалежно від параметрів учасників.

Коли завдань три, а набраний бал $b = 1$, можливі вектори відповідей (1; 0; 0), (0; 1; 0) та (0; 0; 1). Умовна ймовірність того, що у цій ситуації буде дана відповідь на перше питання, аналогічно отримується:

$$P\{(1;0;0) \mid b = 1\} = \frac{\varepsilon_1}{\varepsilon_1 + \varepsilon_2 + \varepsilon_3}.$$

Якщо $b = 2$, можливі варіанти відповідей (1; 1; 0), (1; 0; 1) та (0; 1; 1). Ймовірність того, що буде дана відповідь саме на перше питання, дорівнює сумі двох умовних ймовірностей:

$$P\{(1;1;0)|\,b=2\}+P\{(1;0;1)|\,b=2\}=\frac{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3}{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3}=$$

$$=\frac{\varepsilon_1(\varepsilon_2+\varepsilon_3)}{\varepsilon_1\varepsilon_2+\varepsilon_1\varepsilon_3+\varepsilon_2\varepsilon_3}.$$

У знаменнику маємо попарні (тому, що $b=2$) добутки параметрів усіх завдань, у чисельнику – тільки ті, що містять параметр даного завдання. Знаменник є елементарна симетрична функція другого порядку від змінних $\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\varepsilon_3)$. Позначимо її $\gamma_2(\boldsymbol{\varepsilon})$. Тоді у дужках у чисельнику маємо симетричну функцію першого порядку, з аргументів якої вилучено параметр першого завдання ε_1. Позначимо це $\gamma_1^{(1)}(\boldsymbol{\varepsilon})$. Елементарна симетрична функція нульового порядку дорівнює 1.

Наприклад, умовна ймовірність того, що при набраних трьох балах ($b=3$) з чотирьох завдань з параметрами $\varepsilon_1,\varepsilon_2,\varepsilon_3,\varepsilon_4$, вірна відповідь буде дана саме на друге питання є:

$$\frac{\varepsilon_2(\varepsilon_1\varepsilon_3+\varepsilon_1\varepsilon_4+\varepsilon_3\varepsilon_4)}{\varepsilon_1\varepsilon_2\varepsilon_3+\varepsilon_1\varepsilon_2\varepsilon_4+\varepsilon_1\varepsilon_3\varepsilon_4+\varepsilon_2\varepsilon_3\varepsilon_4}=\frac{\varepsilon_2\cdot\gamma_2^{(2)}(\boldsymbol{\varepsilon})}{\gamma_3(\boldsymbol{\varepsilon})}.$$

Цей результат можна узагальнити. Ймовірність того, що деякий учасник, який набрав b балів, вірно відповість на *j*-те питання, можна записати $\dfrac{\varepsilon_j\cdot\gamma_{b-1}^{(j)}(\boldsymbol{\varepsilon})}{\gamma_b(\boldsymbol{\varepsilon})}$.

Функція умовної вірогідності у цьому випадку дорівнює добутку умовних ймовірностей для всіх осіб $i=\overline{1,n}$ отримати вектор відповідей (a_{ij}) за умови, що набрано b_i балів:

$$L(\boldsymbol{\varepsilon})=\prod_{i=1}^{n}\frac{P\{(a_{ij})|\,\varsigma_i,\boldsymbol{\varepsilon}\}}{P\{b_i|\,\varsigma_i,\boldsymbol{\varepsilon}\}},\ \text{де}$$

$$P\{(a_{ij})|\,\varsigma_i,\boldsymbol{\varepsilon}\}=\prod_{j=1}^{k}\frac{(\varsigma_i\varepsilon_j)^{a_{ij}}}{1+\varsigma_i\varepsilon_j}=\frac{\exp\left(\sum_{j=1}^{k}\theta_i a_{ij}\right)\cdot\prod_{j=1}^{k}\varepsilon_j^{a_{ij}}}{\prod_{j=1}^{k}(1+\varsigma_i\varepsilon_j)}=\frac{\varsigma_i^{b_i}\cdot\prod_{j=1}^{k}\varepsilon_j^{a_{ij}}}{\prod_{j=1}^{k}(1+\varsigma_i\varepsilon_j)},$$

$$P\{b_i \mid \varsigma_i, \boldsymbol{\varepsilon}\} = \sum_{(b_i)} \frac{\varsigma_i^{b_i} \cdot \prod_{j=1}^{k} \varepsilon_j^{a_{ij}}}{\prod_{j=1}^{k}(1+\varsigma_i \varepsilon_j)} = \frac{\varsigma_i^{b_i} \cdot \sum_{(b_i)} \prod_{j=1}^{k} \varepsilon_j^{a_{ij}}}{\prod_{j=1}^{k}(1+\varsigma_i \varepsilon_j)} \ .$$

Тут $\sum_{(b_i)}$ – сума по всіх векторах відповідей, які дають первинний бал b_i, а сума добутків $\sum_{(b_i)} \prod_{j=1}^{k} \varepsilon_j^{a_{ij}}$ є симетрична функція порядку b_i, тобто $\sum_{(b_i)} \prod_{j=1}^{k} \varepsilon_j^{a_{ij}} = \gamma_{b_i}(\boldsymbol{\varepsilon})$. Остаточно функція умовної вірогідності для методу CML має вигляд:

$$L(\boldsymbol{\varepsilon}) = \prod_{i=1}^{n} \frac{\prod_{j=1}^{k} \varepsilon_j^{a_{ij}}}{\gamma_{b_i}(\boldsymbol{\varepsilon})} = \frac{\prod_{j=1}^{k} \left(\varepsilon_j\right)^{\sum_{i=1}^{n} a_{ij}}}{\prod_{i=1}^{n} \gamma_{b_i}(\boldsymbol{\varepsilon})} = \frac{\prod_{j=1}^{k} \left(\varepsilon_j\right)^{c_j}}{\prod_{i=1}^{n} \gamma_{b_i}(\boldsymbol{\varepsilon})}, \qquad (2.47)$$

де $c_j = \sum_{i=1}^{n} a_{ij}$ - первинний бал j-го завдання.

Прирівнюючи до нуля похідні по ε_j від логарифма цієї функції, отримаємо систему k нелінійних рівнянь з k невідомими $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$:

$$c_j = \sum_{b=1}^{k-1} n_b \frac{\varepsilon_j \gamma_{b-1}^{(j)}(\varepsilon)}{\gamma_b(\varepsilon)} = \sum_{b=1}^{k-1} n_b \cdot P\{x_j = 1 \mid b\}, \quad j = \overline{1,k} \ . \qquad (2.48)$$

У цій системі первинний бал завдання дорівнює очікуваному, який обчислюється як сума балів за це завдання по кожній групі учасників кількості n_b, які набрали b балів $(0 < b < k)$. Тобто, отримали систему методу моментів. Для даної системи теж потрібно використовувати ітераційні процедури. Або доручити це виконати якому-небудь математичному пакету. Далі будують оцінки рівнів підготовленості учасників за відомими параметрами завдань. Починати побудову оцінок за CML процедурою важливо з

видалення рядків та стовпців матриці відповідей, що містять лише самі 0 або самі 1. На відміну від CML процедуру JML ще називають безумовною (unconditional).

Крім розглянутих вище різновидів методів максимальної вірогідності розроблені й інші методики оцінки латентних параметрів. Це, у першу чергу, методи, засновані на Байєсівському підході (EAP – Expected a Posteriri estimator), на властивостях ланцюгів Маркова та методу Монте-Карло (MCMC – Markov Chain Monte Carlo), на нелінійному факторному аналізі та на комбінуванні різних підходів та методик.

2.5. Алгоритм PROX розрахунку оцінок параметрів

Розглянемо метод PROX побудови оцінок латентних параметрів θ і δ, який не вимагає ітераційних процедур і може використовуватись при обчисленнях вручну. Він також не потребує великих об'ємів вибірок. Його ще називають методом нормальної апроксимації (Normal Approximation algorithm), оскільки він будується на припущеннях про нормальність розподілів емпіричних даних тестування та значень латентних змінних.

Нехай рівні підготовки учасників мають нормальний розподіл з середнім *M* та стандартним відхиленням σ: $\theta_i \in N(M,\sigma^2)$; а складності завдань – нормальний з параметрами відповідно $\delta_j \in N(H,\omega^2)$. Початкові значення оцінюваних параметрів вибирають, враховуючи рівність (1.14), де оцінками ймовірностей будуть відповідні частоти. Наприклад, частка правильних відповідей (успіхів) для *i*-го учасника становить $p_i = \frac{b_i}{k}$, тоді частка неуспіхів $q_i = 1 - p_i$. Початковою оцінкою логітів рівня підготовки *i*-го учасника, який відповідає на завдання нульової складності (*H*=0), будемо вважати $\theta_i^0 = \ln\left(\frac{p_i}{q_i}\right) = \ln\left(\frac{b_i}{k-b_i}\right)$. Аналогічно, початкове значення складності завдань $\delta_j^0 = \ln\left(\frac{q_j}{p_j}\right) = \ln\left(\frac{n-c_j}{c_j}\right)$, де $p_j = \frac{c_j}{n}$

– частка успіхів у *j*-му завданні, $q_j = 1 - p_j$ – частка неуспіхів. Початкові оцінки матимуть різні середні та дисперсії, щоб звести їх до єдиної інтервальної шкали, покладемо

$$\hat{\theta}_i = H + X \ln\left(\frac{b_i}{k - b_i}\right),\ \hat{\delta}_j = M + Y \ln\left(\frac{n - c_j}{c_j}\right), \qquad (2.49)$$

де уточнюючі множники *X* та *Y* вибираються, виходячи з властивостей нормального розподілу, так, щоб вирівнялись дисперсії оцінок. Якщо *V* – дисперсія початкових значень оцінок рівнів підготовленості θ_i^0, а *U* – дисперсія початкових значень δ_j^0, то

$$X = \sqrt{\frac{1 + U/2{,}89}{1 - UV/8{,}35}}, \qquad Y = \sqrt{\frac{1 + V/2{,}89}{1 - UV/8{,}35}}. \qquad (2.50)$$

Тут множники $2{,}89 = 1{,}7^2$, $8{,}35 = 1{,}7^4$ з'являються через заміну нормальної кривої логістичною.

Розглянемо алгоритм методу на конкретному прикладі тестування [19], де у матриці відповідей після вилучення рядків та стовпців, що складаються з одних лише нулів або одиниць, залишилось 34 рядки (учасники) та 14 стовпців (питань). Підраховані первинні бали завдань c_j та частоти появи цих балів f_j (таблиця 5). Тут уже немає тих завдань, бали яких 34 або 0, бал 30 зустрічається для двох завдань, 1 бал є у трьох завдань, $\sum_j f_j = 14$. У таблиці 6 подано згруповані дані по учасниках тестування: b_i – первинний бал учасника, n_{b_i} – кількість учасників з балом b_i, $\sum_{b_i} n_{b_i} = 34$.

Далі послідовність дій така:

1. Підраховуємо частки правильних відповідей на кожне питання $p_j = \frac{c_j}{n}$ ($j = \overline{1,k}$) та обчислюємо початкові оцінки складностей завдань $\delta_j^0 = \ln\left(\frac{1 - p_j}{p_j}\right)$ (таблиця 5).

2. Знаходимо середнє значення початкових оцінок складностей завдань, враховуючи частоти, $\overline{\delta} = \frac{1}{k}\sum_j f_j \cdot \delta_j^0 = \frac{2{,}62}{14} = 0{,}19$.

3. Знаходимо дисперсію *U* по множині δ_j^0:

$$U = \frac{1}{k-1}\left[\sum_j f_j \cdot \left(\delta_j^0\right)^2 - \sum_j f_j \cdot \left(\overline{\delta}\right)^2\right] = \frac{76{,}00 - 0{,}49}{13} = 5{,}81.$$

4. Підраховуємо частки правильних відповідей кожного учасника на всі питання тесту $p_i = \frac{b_i}{k}$ $(i = \overline{1,n})$ та обчислюємо початкові оцінки рівнів підготовленості $\theta_i^0 = \ln\left(\frac{p_i}{1-p_i}\right)$ (таблиця 6).

Таблиця 5.

c_j	f_j	$p_j = \frac{c_j}{n}$	δ_j^0	$f_j\delta_j^0$	$f_j\left(\delta_j^0\right)^2$	$f_j\left(\overline{\delta}\right)^2$	$\delta_j^0 - \overline{\delta}$	$\hat{\delta}_j$
32	1	0,94	-2,77	-2,77	7,69	0,04	-2,96	-3,87
31	2	0,91	-2,34	-4,67	10,91	0,07	-2,52	-3,29
30	2	0,88	-2,01	-4,03	8,12	0,07	-2,20	-2,88
27	1	0,79	-1,35	-1,35	1,82	0,04	-1,54	-2,01
24	1	0,71	-0,88	-0,88	0,77	0,04	-1,06	-1,39
12	1	0,35	0,61	0,61	0,37	0,04	0,42	0,55
7	1	0,21	1,35	1,35	1,82	0,04	1,16	1,52
6	1	0,18	1,54	1,54	2,37	0,04	1,35	1,77
3	1	0,09	2,34	2,34	5,45	0,04	2,15	2,81
1	3	0,03	3,50	10,49	36,68	0,11	3,31	4,32
Суми				2,62	76,00	0,49		

5. Знаходимо середнє значення початкових оцінок рівнів підготовленості, враховуючи частоти, $\overline{\theta} = \frac{1}{n}\sum_i n_{b_i}\theta_i^0 = \frac{-0{,}96}{34} = -0{,}03$.

6. Знаходимо дисперсію V по множині θ_i^0:

$$V = \frac{1}{n-1}\left[\sum_i n_{b_i}\cdot\left(\theta_i^0\right)^2 - \sum_i n_{b_i}\cdot\left(\overline{\theta}\right)^2\right] = \frac{15{,}22-0{,}03}{33} = 0{,}46.$$

7. Обчислюємо уточнюючі множники:

$$X = \sqrt{\frac{1+U/2{,}89}{1-UV/8{,}35}} = \sqrt{\frac{1+5{,}81/2{,}89}{1-5{,}81\cdot 0{,}46/8{,}35}} = 2{,}10,$$

$$Y = \sqrt{\frac{1+V/2{,}89}{1-UV/8{,}35}} = \sqrt{\frac{1+0{,}46/2{,}89}{1-5{,}81\cdot 0{,}46/8{,}35}} = 1{,}31.$$

Таблиця 6.

b_i	n_{b_i}	$p_i = \frac{b_i}{k}$	θ_i^0	$n_b\theta_i^0$	$n_b\left(\theta_i^0\right)^2$	$n_b\left(\overline{\theta}\right)^2$	$\hat{\theta}_i$
1	0	0,07	-2,56	0,00	0,00	0,00	-5,40
2	1	0,14	-1,79	-1,79	3,21	0,00	-3,77
3	2	0,21	-1,30	-2,60	3,38	0,00	-2,73
4	2	0,29	-0,92	-1,83	1,68	0,00	-1,93
5	2	0,36	-0,59	-1,18	0,69	0,00	-1,24
6	3	0,43	-0,29	-0,86	0,25	0,00	-0,61
7	12	0,50	0,00	0,00	0,00	0,01	0,00
8	5	0,57	0,29	1,44	0,41	0,00	0,61
9	4	0,64	0,59	2,35	1,38	0,00	1,24
10	1	0,71	0,92	0,92	0,84	0,00	1,93
11	2	0,79	1,30	2,60	3,38	0,00	2,73
12	0	0,86	1,79	0,00	0,00	0,00	3,77
13	0	0,93	2,56	0,00	0,00	0,00	5,40
Суми				-0,96	15,22	0,03	

8. Обчислюємо відхилення $\delta_j^0 - \overline{\delta}$, щоб початок відліку на єдиній інтервальній шкалі збігався з $H = 0$. Можна показати, що у цьому випадку $M \approx -Y\overline{\delta}$, тому остаточно маємо оцінки

$\hat{\delta}_j = Y \cdot \left(\delta_j^0 - \overline{\delta}\right)$, $\hat{\theta} = X \cdot \theta_i^0$, які обчислено у останніх стовпчиках таблиць 5 та 6. Тут маємо $\sum_{j=1}^{14} f_j \hat{\delta}_j = 0$, тобто тест збалансований за складністю.

9. Оцінюємо стандартні похибки вимірювання $SE(\hat{\theta}_i)$ та $SE(\hat{\delta}_j)$, які у даному випадку розраховується за формулами:

$$SE(\hat{\theta}_i) = \frac{X}{\sqrt{p_i(k-b_i)}} = \frac{X}{\sqrt{kp_i(1-p_i)}} = \frac{X}{\sqrt{kp_i q_i}},$$

$$SE(\hat{\delta}_j) = \frac{Y}{\sqrt{p_j(n-c_j)}} = \frac{Y}{\sqrt{np_j(1-p_j)}} = \frac{Y}{\sqrt{np_j q_j}}.$$

Метод PROX використовується у деяких програмних продуктах для побудови початкового наближення методу максимальної вірогідності з метою зменшення кількості необхідних ітерацій.

3. ОПИСОВІ ФУНКЦІЇ ТЕСТУ

3.1. Характеристична функція тесту

Характеристична функція тесту вказує на очікуваний сумарний бал для всіх рівнів підготовки θ і обчислюється як сума очікуваних балів за кожне завдання:

$$E(\theta) = \sum_{j=1}^{k} E_j(\theta) = \sum_{j=1}^{k} \sum_{l=0}^{m_j} l \cdot p_{lij}. \qquad (3.1)$$

Для тесту, який складається лише з дихотомічних завдань, матимемо суму ймовірностей вірного виконання кожного завдання $E(\theta) = \sum_{j=1}^{k} E_j(\theta) = \sum_{j=1}^{k} P_j(\theta)$. У цьому випадку суму часто ділять на максимально можливий бал за тест, щоб отримати для кожного θ очікувану пропорцію виконаних завдань. Це зручно, коли порівнюються тести з різною кількістю завдань. Графік характеристичної функції називають *характеристичною кривою тесту* ТСС (Test Characteristic Curve). На рис.19 *а*) характеристичні криві двох тестів, які дають залежність очікуваного бала від рівня підготовле-

ності. В одному (суцільна крива) – три дихотомічні завдання, у іншому (пунктирна крива) – три дихотомічні та одне політомічне з максимальною кількістю 2 бали. На рис. 19 *b*) ТСС для дихотомічного тесту, виражена у відношеннях очікуваного бала до максимально можливого бала 3.

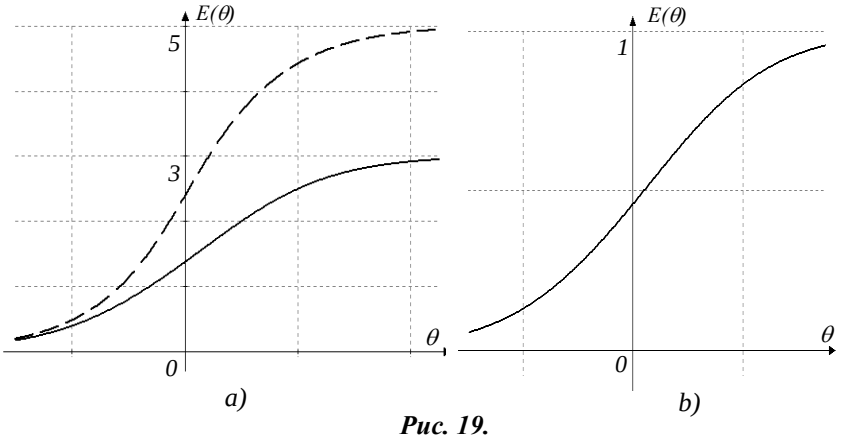


Рис. 19.

Якщо розподіл учасників тестування за рівнем підготовленості θ відомий, то ТСС визначає форму розподілу підсумкових балів. Неприпустими є горизонтальні ділянки ТСС або її значна пологість, бо тоді тест не розрізняє учасників з різними рівнями підготовленості.

3.2. Інформаційна функція завдань та тесту

Термін «інформація» в сучасній теорії тестування використовується як статистичний індикатор якості оцінювання латентних параметрів, а саме параметра рівня підготовки опитаних θ. Оскільки відповіді учасників тестування на завдання тесту надають інформацію для кожного значення на шкалі θ, то інформація уявляється швидше як функція, а не окреме число. Формула для інформації може бути отримана різними способами.

Так, Birnbaum (1968) показав, що величина інформації пов'язана з довжиною довірчого інтервалу навколо оцінки $\hat{\theta}$, Kendall та Stuart (1961) вказали, що інформація пов'язана з асимп-

тотичною стандартною похибкою оцінок $\hat{\theta}$, отриманих з використанням методу максимальної вірогідності, Lord (1980), Hambleton та Swaminathan (1985) означали інформацію, як індикатор властивості первинних балів відображати реальну різницю між рівнями підготовки, тобто, як індикатор точності, з якою θ може бути оцінена у даному діапазоні рівнів підготовки.

Припустимо, що оцінку $\hat{\theta}$ реального рівня підготовленості θ було здійснено методом максимальної вірогідності у рамках одновимірної моделі, яка добре моделює реальні дані. Оцінка $\hat{\theta}$ є випадковою величиною, оскільки при повторних тестуваннях може приймати різні значення. Нехай $M\left\{\hat{\theta}|\theta_i\right\}$ та $\sigma\left\{\hat{\theta}|\theta_i\right\}$ – математичне сподівання та стандартне відхилення оцінки для особи з рівнем θ_i. Якщо дві особи з рівнями підготовки θ_1 та θ_2 оцінюються одним тестом, то можливою мірою чутливості тесту (його здатності відрізнити ці рівні) є величина $\dfrac{M\left\{\hat{\theta}|\theta_2\right\}-M\left\{\hat{\theta}|\theta_1\right\}}{\sigma\left\{\hat{\theta}|\theta_1\theta_2\right\}}$ – відстань між середніми у одиницях стандартного відхилення, де $\sigma\left\{\hat{\theta}|\theta_1\theta_2\right\}$ – об'єднане стандартне відхилення обох θ. Ця величина залежить від відстані між θ_1 та θ_2, чим більша відстань, тим більша кількість одиниць стандартного відхилення між ними. Щоб досліджувати чутливість тесту у різних точках осі θ, знайдемо, яка кількість одиниць стандартного відхилення відповідає одиниці масштабу на θ-осі: $\dfrac{M\left\{\hat{\theta}|\theta_2\right\}-M\left\{\hat{\theta}|\theta_1\right\}}{\sigma\left\{\hat{\theta}|\theta_1\theta_2\right\}\cdot(\theta_2-\theta_1)}$. Чутливість у точці отримаємо, переходячи до границі при $\theta_2\to\theta_1$, тоді об'єднане стандартне відхилення $\sigma\left\{\hat{\theta}|\theta_1\theta_2\right\}$ наближається до стандартного відхилення у одній точці. Отже,

$$\lim_{\theta_2\to\theta_1}\frac{M\left\{\hat{\theta}|\theta_2\right\}-M\left\{\hat{\theta}|\theta_1\right\}}{\sigma\left\{\hat{\theta}|\theta_1\theta_2\right\}\cdot(\theta_2-\theta_1)}=\frac{\partial M\left\{\hat{\theta}|\theta\right\}}{\partial\theta}\cdot\frac{1}{\sigma\left\{\hat{\theta}|\theta\right\}}. \qquad (3.2)$$

У (3.2) маємо функцію від θ, яка показує, наскільки добре

відмінності в θ можуть бути виявлені за допомогою тестових завдань, на основі яких отримані оцінки $\hat{\theta}$.

Lord (1980) показав, що при побудові оцінок методом максимальної вірогідності квадрат величини, оберненої до (3.2), дорівнює асимптотичній дисперсії цієї оцінки.

Квадрат виразу (3.2) називають кількістю інформації, яка забезпечується оцінкою $\hat{\theta}$ про реальний рівень підготовки θ, а саму залежність називають *інформаційною функцією*:

$$I(\theta) = \frac{\left(E'(\theta)\right)^2}{\sigma^2(\theta)}. \tag{3.3}$$

Тоді попередній висновок набуде вигляду $SE(\hat{\theta}) = \frac{1}{\sqrt{I(\theta)}}$.

Якщо завдання дихотомічне, то його інформаційну функцію можна записати

$$I_j(\theta) = \frac{\left(P_j'(\theta)\right)^2}{P_j(\theta) \cdot (1 - P_j(\theta))}, \tag{3.4}$$

де $P_j(\theta)$ – ймовірність правильної відповіді на *j*-те питання, яка залежить від обраної моделі. У чисельнику функції (3.4) маємо похідну характеристичної функції завдання, яка досягає максимуму у точці перегину, там, де дотична утворює з віссю θ найбільший кут. Оскільки точці перегину відповідає значення $\theta = \delta_j$, то можемо зробити важливий висновок, що для вимірювання конкретного значення θ_i найбільш інформативними будуть завдання зі складністю $\delta \approx \theta_i$, тобто з околу точки θ_i осі θ. Чим менша відстань, на якій знаходяться завдання від θ_i, тим меншою є стандартна похибка вимірювання даного значення θ_i.

Для однопараметричної моделі Раша, враховуючи вираз (1.11), маємо $P_j'(\theta) = P_j(\theta) \cdot Q_j(\theta)$, тому

$$I_j(\theta) = P_j(\theta) \cdot Q_j(\theta). \tag{3.5}$$

Для двопараметричної моделі 2PL відповідно до (1.16) маємо $P_j'(\theta) = Dd_j \cdot P_j(\theta) \cdot Q_j(\theta)$, тому

$$I_j(\theta) = \left(Dd_j\right)^2 \cdot P_j(\theta) \cdot Q_j(\theta). \qquad (3.6)$$

Вираз інформаційної функції завдання у трипараметричній моделі 3PL у різних джерелах можна зустріти у різних варіантах:

$$I_j(\theta) = \frac{\left(Dd_j\right)^2 \cdot \left(1 - c_j\right)}{\left(c_j + \exp(Dd_j(\theta - \delta_j)\right) \cdot \left(1 + \exp(-Dd_j(\theta - \delta_j)\right)^2}, \qquad (3.7^*)$$

$$I_j(\theta) = \left(Dd_j\right)^2 \frac{Q_j(\theta)}{P_j(\theta)} \cdot \left(\frac{P_j(\theta) - c_j}{1 - c_j}\right)^2, \qquad (3.7^{**})$$

$$I_j(\theta) = \left(Dd_j\right)^2 \cdot p_j(\theta) \cdot \left(1 - p_j(\theta)\right) \cdot \left(1 - \frac{c_j}{P_j(\theta)}\right). \qquad (3.7^{***})$$

Кожну з цих формул нескладно отримати, якщо для $P_j(\theta)$ використати вираз (1.17), а ймовірність правильної відповіді без угадування $p_j(\theta)$ знаходити за формулою (1.16).

З аналізу (3.5) випливає, що максимум інформаційної функції завдання у моделі Раша дорівнює 0,25 і досягається при $P_j(\theta) = 0{,}5$, тобто у точці $\theta = \delta_j$, де характеристична функція завдання має перегин. Для моделі 2PL, аналізуючи (3.6), отримуємо аналогічний результат, але максимальне значення $0{,}25\left(Dd_j\right)^2$. Тут роздільна здатність завдання суттєво впливає на точність оцінюваних параметрів. З одного боку добре, що збільшується точність оцінки в околі оцінюваного параметра, але погано, що може звужуватися робоча область завдання. Якщо в однопараметричній моделі кожне завдання задовільно перекриває інтервал $\delta_j \pm 1$, то у двопараметричній моделі для цього діапазону можуть знадобитись два або більше завдань, залежно від d_j. Іншими словами, збільшення роздільної здатності завдання призводить до зменшення числа учасників тестування, на яких воно розраховане. На рис.20 зображена характеристична (*I*) та інформаційна (*II*) функції деякого завдання у 2PL при $Dd_j = 2{,}89$.

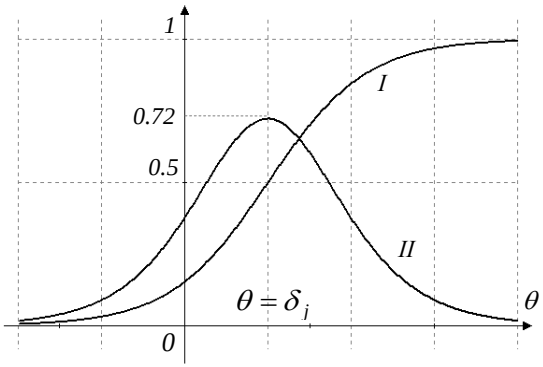


Рис. 20.

У трипараметричній моделі параметр псевдоугадування c_j помітно знижує точність оцінок латентних параметрів та уповільнює збіжність ітераційних процедур, що використовуються для оцінювання. Досліджуючи (3.7*), отримаємо, що екстремум цієї функції досягається у точці

$$\theta_{\max} = \delta_j + \frac{1}{Dd_j} \ln\left[0{,}5\left(1+\sqrt{1+8c_j}\right)\right]. \qquad (3.8)$$

Якщо угадування мінімальне $c_j = 0$, то $\theta_{\max} = \delta_j$. Загалом при $c_j > 0$ завдання забезпечує максимальну інформацію на рівні здібностей вищому, ніж його складність. Обчислення значень $I(\theta_{\max})$ для різних c_j вказує на зростання кількості інформації при зменшенні c_j і максимального значення $I(\theta_{\max})$ досягає при $c_j = 0$.

На рис.21 зображено інформаційні функції шести різних завдань з параметрами:

j	δ_j	d_j	c_j
1	*1.0*	*1.8*	*0*
2	1.0	0.8	0
3	1.0	1.8	0.25
4	-1.5	1.8	0
5	-0.5	1.2	0.1
6	0.5	0.4	0.15

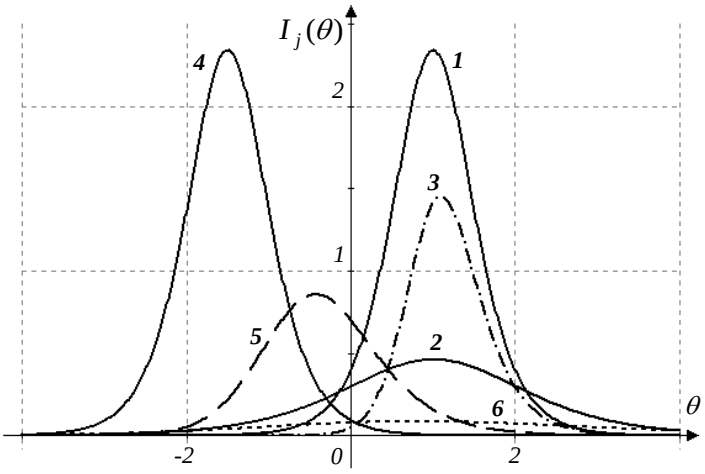


Рис. 21.

Завдання 1 і 2 мають однакову складність $\delta = 1$, але більш інформативним (дає меншу похибку вимірювання) на проміжку між 0 та 2 є завдання 1, у якого коефіцієнт дискримінуючої здатності вищий. Однак за межами цього проміжку з них двох більш ефективним може виявитись 2. Завдання 1 і 3 відрізняються лише параметром угадування. Його наявність зменшує максимум інформації, який зміщується у сторону вищих рівнів підготовки. Завдання 3 забезпечує меншу похибку вимірювання для сильніших опитаних. Завдання 1 та 4 мають лише різні складності, які відрізняються так сильно, що обидва завдання в області малих від'ємних θ не спроможні забезпечити належний рівень точності. Там краще працює завдання 5. Низька диференціююча здатність та наявність параметра угадування у завданні 6 приводить до того, що воно не придатне для оцінювання параметра θ всіх рівнів підготовки.

Завдяки властивості адитивності інформація, отримана при вимірюванні даного θ за допомогою всього тесту, є сумою окремих складових $I_j(\theta)$ для кожного завдання:

$$I(\theta) = \sum_{j=1}^{k} I_j(\theta), \qquad (3.9)$$

де $I(\theta)$ – *інформаційна функція тесту*. Hambleton та Swaminathan (1985) отримали, що інформація тесту в кожній точці θ дорівнює від'ємному математичному сподіванню другої похідної логарифма функції максимальної вірогідності, тобто $I(\theta) = -M\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right]$.

Для тесту з дихотомічними завданнями інформаційну функцію зручно записувати у вигляді:

$$I(\theta) = \sum_{j=1}^{k} \frac{\left(P_j'(\theta)\right)^2}{P_j(\theta) \cdot (1 - P_j(\theta))}.$$

На рис.22 для прикладу зображено інформаційні функції двох тестів (криві 4), у кожному з яких по 3 завдання. У першому тесті (*a*) завдання мають рівномірно розподілену складність вздовж осі і інформаційна функція має один яскраво виражений екстремум. Інформаційна функція другого тесту (*b*) має «провал», оскільки тест не однорідний, одне завдання має значно більшу складність, ніж два інші. Ситуацію можна змінити, якщо додати завдання проміжної складності, або змінити складність існуючих завдань.

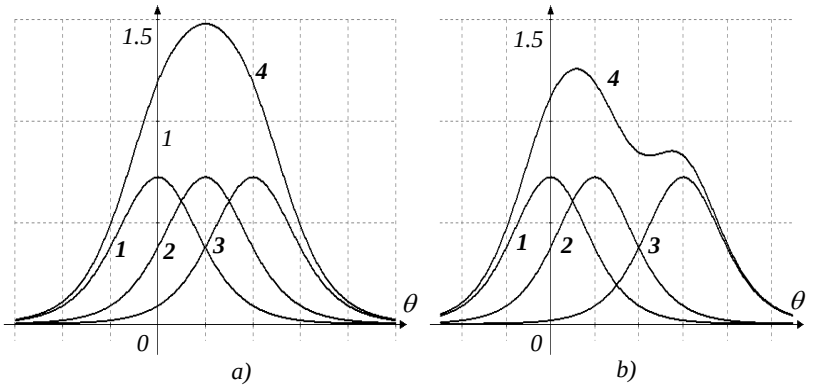


Рис. 22.

Інформаційна функція гарного збалансованого тесту повинна мати один чітко виражений екстремум. Якщо графік інформаційної функції має пологий, не чітко виражений екстремум, говорять про зниження ефективності усього тесту. У випадку, коли є

кілька локальних екстремумів, наприклад, два при θ_1 та θ_2, тест потребує вдосконалення. Якщо кількість завдань у тесті не велика, то потрібно додавати завдання, які мають проміжну складність $\theta_1 < \delta < \theta_2$, щоб ліквідувати «провали» між сусідніми екстремумами. Якщо кількість завдань у тесті досить велика ($k > 100$), то його варто розбити на два тести, один з яких буде ефективним для вибірки з середнім значенням θ поблизу θ_1, а інший – для іншої вибірки з $\theta \approx \theta_2$.

Для політомічних завдань інформаційна функція дорівнює не простій, а зваженій сумі інформації, яку вносить кожна категорія. Samejima (1969) для інформації завдання отримала формулу

$$I_j(\theta) = \sum_{l=0}^{m_j} \frac{\left(P'_{lj}(\theta)\right)^2}{P_{lj}(\theta)}, \qquad (3.10)$$

де $P_{lj}(\theta)$ – ймовірність того, що особа з рівнем підготовленості θ у *j*-му завданні подолає *l* кроків, вигляд якої залежить від обраної моделі. Наприклад, у моделі Partial Credit можна отримати такий вираз для інформаційної функції завдання:

$$I_j(\theta) = \sum_{l=0}^{m_j} \left[l - \sum_{k=0}^{m_j} k \cdot P_{kj}(\theta) \right]^2 \cdot P_{lj}(\theta). \qquad (3.11)$$

Інформаційна функція тесту тут також дорівнює сумі інформаційних функцій завдань.

На рис.23 зображено інформаційні функції трьох завдань у моделі Partial Credit, кожне з яких має по три кроки з різними складностями. Перше завдання має перший крок легший $\delta_{11} = -2$, ніж другий $\delta_{12} = 2$. У другому завданні, навпаки $\delta_{21} = 2$, $\delta_{22} = -2$. Третє завдання має два кроки однакової складності $\delta_{31} = \delta_{32} = 0$.

Більш інформативним і придатним для оцінювання рівнів підготовки з околу $\theta = 0$ є завдання 2, у якого складнішим є перший крок. При оцінюванні рівнів $\theta > 2$ та $\theta < -2$ краще працює завдання 1, у якого навпаки, перший крок легший. Завдання 3 з кроками однакової складності у цілому менш інформативне. Також очевидно, що з двох завдань більш інформативним є те, у якого відстань між складностями категорій менша.

Інформаційні криві політомічних завдань можуть дуже відрізнятися за формою від відповідних кривих дихотомічних завдань.

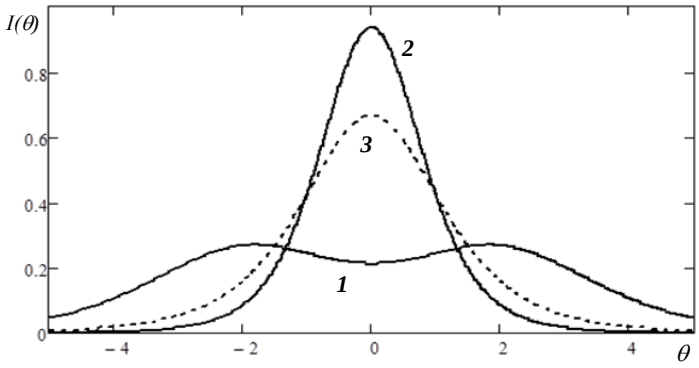


Рис. 23.

Відомий також факт (Samejima, 1969), що збільшення кількості категорій у завданні веде до підвищення точності оцінювання. Ілюструє його рисунок 24. Перше завдання має усього два кроки, складності яких дуже сильно відрізняються $\delta_{11} = -2$, $\delta_{12} = 2$; у другому завданні додано проміжний крок так, що $\delta_{21} = -2$, $\delta_{22} = 0$, $\delta_{23} = 2$; третє завдання має п'ять кроків $\delta_{31} = -2$, $\delta_{32} = -1$, $\delta_{33} = 0$, $\delta_{34} = 1$ і $\delta_{35} = 2$.

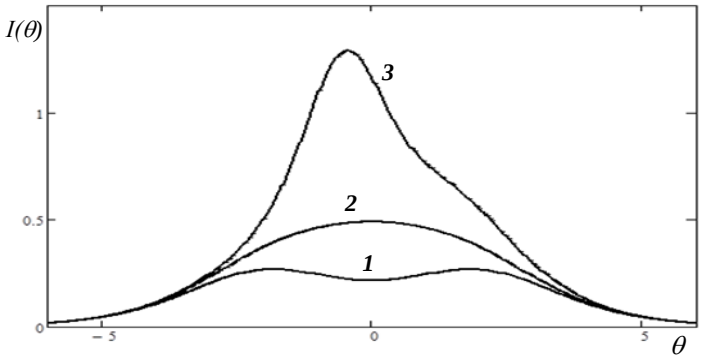


Рис. 24.

Очевидне підвищення інформативності завдань із збільшенням кількості категорій у досить широкому діапазоні θ.

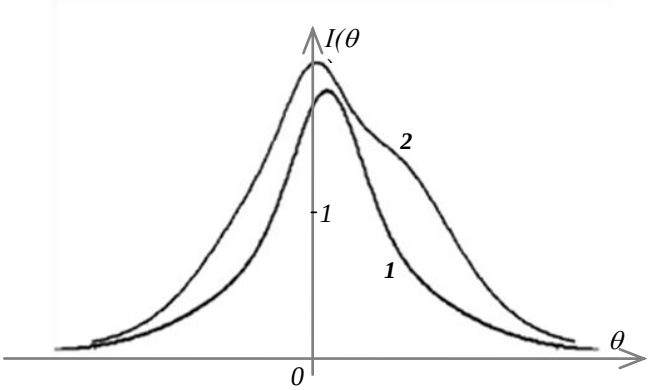


Рис. 25.

Інформаційні функції тестів з рис.23 (крива 1) та 24 (крива 2) зображені на рис.25. Зауваження щодо їх форми такі ж, як у випадку дихотомічних завдань.

Зауваження 1. Інформаційну функцію можна використовувати для знаходження довірчого інтервалу для параметра θ. Якщо припустити, що $\hat{\theta}$ розподілена нормально та згадати, що $SE(\hat{\theta}) = 1/\sqrt{I(\hat{\theta})}$, то довірчий інтервал з надійністю, наприклад, 95% матиме вигляд $\hat{\theta} - \frac{1.96}{\sqrt{I(\hat{\theta})}} < \theta < \hat{\theta} + \frac{1.96}{\sqrt{I(\hat{\theta})}}$.

3.3. Функція відносної ефективності

Можна показати, що відношення інформаційних функцій двох завдань має властивість інваріантності. Тобто, при будь-якому перетворенні θ до θ^* матимемо $\frac{I_i(\theta)}{I_j(\theta)} = \frac{I_i(\theta^*)}{I_j(\theta^*)}$. Це дає підставу для порівняння ефективності різних тестів. Припустимо, що два тести *X* та *Y* оцінюють одну і ту ж приховану характеристику θ. Ефективність тесту *Y* по відношенню до тесту *X* характеризують відношенням інформаційних функцій цих тестів у відпо-

відних точках осі θ:

$$E(Y,X)=\frac{I(\theta,Y)}{I(\theta,X)}. \qquad (3.12)$$

Функцію (3.12) називають *функцією відносної ефективності*. Вона також має властивість інваріантності при зміні масштабу.

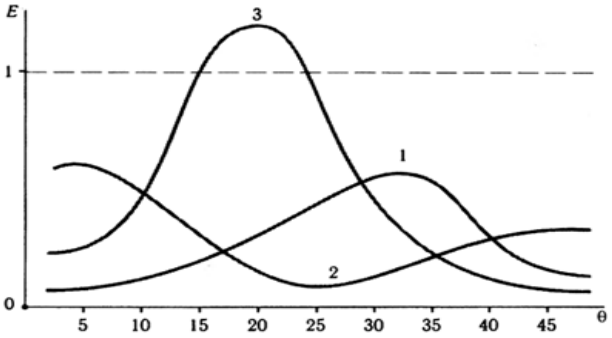


Рис. 26.

Функцію відносної ефективності зручно використовувати при моделюванні тестів із різними наперед заданими властивостями без попереднього збору емпіричної інформації при наявності банку відкаліброваних завдань. На рис.26 представлено приклад моделювання трьох таких тестів Y_1, Y_2 та Y_3, які утворені або з даного тесту *X*, який складався із 50 завдань, або додаванням нових завдань відомої складності із банку завдань. На горизонтальній осі відмічено бали учасників тестування, ефективність $E=1$ відповідає початковому тесту.

Тест Y_1 утворений з початкового відбором 25 найскладніших завдань тесту *X*. Такий тест виявився б менш ефективним для всіх опитаних (крива 1 знаходиться нижче горизонтальної лінії $E=1$ для всіх θ). У другий тест Y_2 відібрано 25 найлегших завдань з 50 завдань початкового тесту. Такий тест демонструє суттєве зменшення ефективності при тестуванні добре підготовлених учасників у порівнянні з попереднім тестом. Навпаки, для опитаних з низькими балами другий тест виявився б більш ефективним. Такого результату слід було очікувати, оскільки угадування слабкими

учасниками відповідей на складні питання значно знижує інформативність тесту. Третій тест Y_3 сформовано вибором із банку 50 нових завдань середньої складності. Такий тест більш ефективний, ніж даний, на вибірці опитаних із середнім підсумковим балом.

Отже, використовуючи функцію відносної ефективності можна змоделювати тест, підбираючи із банку відкалібровані завдання так, щоб тест забезпечував максимальну ефективність на конкретній вибірці учасників тестування.

4. ВІДПОВІДНІСТЬ ЕМПІРИЧНИХ ДАНИХ МОДЕЛІ

4.1. Аналіз залишків

Після збору даних та їх обробки постає питання інтерпретації результатів. Хоча математичні моделі мають багато хороших властивостей, нема гарантії, що зібрані дані повністю відповідають побудованій моделі. Серед дослідників у природничій сфері (наприклад, фізика) переважає такий підхід до моделювання: якщо теоретична залежність не відповідає спостережуваній в експерименті, то роблять висновок, що теорія недостатньо розвинута і намагаються її вдосконалити. Тут такий підхід повністю виправданий, оскільки закони природи не залежать від дослідника, їх не можна виправити. У теорії педагогічних вимірювань можливий інший підхід, оскільки тести повністю знаходяться в руках дослідника. Г.Раш вважав, що якщо емпіричні дані суперечать його теорії, то такі дані потрібно відкидати як недостовірні. А вдосконалювати потрібно не теорію, а дані (частіше тест). Хоча формально модель Раша є однопараметричною теорією IRT, але по суті це окрема теорія, оскільки введення додаткових параметрів у дво- та трипараметричній моделі відбувалося якраз з метою вдосконалення теорії, щоб вона краще описувала емпіричні дані.

Через те, що моделі сімейства Раша створювалися з метою побудови «ідеальних» тестів, далі розглянемо деякі аспекти дослідження відповідності (model fit) емпіричних даних саме для цих моделей. Як правило, таке дослідження включає два основні напрямки, реалізовані у більшості програмних продуктів з аналізу результатів тестування:

- аналіз залишків для виявлення серйозності відхилень;
- статистична перевірка гіпотез про відповідність моделі.

Wright (1977) запропонував кілька зручних статистик для аналізу відповідності даних моделі Раша (fit statistics) на основі стандартизованих залишків, або ж залишкових статистик.

Нехай x_{ij} – спостережувана кількість первинних балів *i*-го учасника за *j*-те завдання; $M(x_{ij})$, $D(x_{ij})$ – математичне сподівання (очікувана кількість балів) та дисперсія первинних балів. Тоді *стандартизованим залишком* називають величину:

$$z_{ij} = \frac{x_{ij} - M(x_{ij})}{\sqrt{D(x_{ij})}}. \qquad (4.1)$$

Для дихотомічних завдань відомо, що $M(x_{ij}) = P_{ij}$, $D(x_{ij}) = P_{ij}(1 - P_{ij})$, де P_{ij} – ймовірність правильної відповіді *i*-го учасника на *j*-те завдання. Якщо завдання політомічне, то очевидно, що $M(x_{ij}) = \sum_{l=0}^{m_j} l \cdot P_{lij}$, а $D(x_{ij}) = \sum_{l=0}^{m_j} \left(l - M(x_{ij})\right)^2 \cdot P_{lij}$. Залишки дозволяють привернути увагу до проблемних завдань/учасників, які ще називаються «викидами», коли сильні учасники дають невірну відповідь, а слабкі – навпаки вірну. Наприклад, якщо для деякого учасника ймовірність правильної відповіді 0.8, а він дає невірну відповідь на дихотомічне питання, то залишок такого неспо-діваного результату $z_{ij} = \frac{0 - 0.8}{\sqrt{0.8 \cdot 0.2}} = -2$. Якщо ж навпаки, ймовірність вірної відповіді мала 0.2, а учасник дає вірну відповідь, то $z_{ij} = 2$. Для прогнозованих відповідей, узгоджених із здоровим глуздом, залишки повинні бути в околі нуля ($-2 < z_{ij} < 2$).

Сума квадратів залишків по всіх учасниках ($i = \overline{1,n}$) може слугувати статистичним індексом завдання, а якщо сумувати по завданнях ($j = \overline{1,k}$), отримаємо характеристику учасника. Далі досліджуватимемо якість завдань, хоча все так само виглядатиме і для учасників.

Незваженою середньо-квадратичною статистикою (mean-square), яка в іноземній літературі ще має назву **outfit**, називають величину

$$U_j^{(1)} = \frac{1}{n}\sum_{i=1}^{n} z_{ij}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(x_{ij} - M(x_{ij})\right)^2}{D(x_{ij})}. \qquad (4.2)$$

Ця статистика має розподіл χ^2 з 1 ступенем волі, її математичне сподівання дорівнює 1, а середньоквадратичне відхилення

$$\sigma\left(U_j^{(1)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{D(x_{ij})} - 4n\right]^{\frac{1}{2}}. \qquad (4.3)$$

Статистика **outfit** має кілька недоліків. Вона дуже чутлива до «викидів», а її критичні точки неможливо визначити однозначно, оскільки вони залежать від об'єму вибірки та від кількості інформації. Щоб знизити вплив екстремальних значень, розглядають *зважену середньо-квадратичну статистику*, яка ще має назву **infit.** Вона враховує кількість інформації від кожного доданка в чисельнику і зменшує вплив екстремальних значень через наявність інформації у знаменнику:

$$U_j^{(2)} = \frac{\sum_{i=1}^{n} z_{ij}^2 D(x_{ij})}{\sum_{i=1}^{n} D(x_{ij})} = \frac{\sum_{i=1}^{n}\left(x_{ij} - M(x_{ij})\right)^2}{\sum_{i=1}^{n} D(x_{ij})}. \qquad (4.4)$$

Ця статистика має теж розподіл χ^2 з 1 ступенем волі, її математичне сподівання дорівнює 1, а середньоквадратичне відхилення

$$\sigma\left(U_j^{(2)}\right) = \frac{\left[\sum_{i=1}^{n} D(x_{ij}) - 4\sum_{i=1}^{n} D^2(x_{ij})\right]^{1/2}}{\sum_{i=1}^{n} D(x_{ij})}. \qquad (4.5)$$

Якщо складність завдання близька до рівня підготовки, кількість інформації велика, вага такого доданка більша. Навпаки, якщо складність завдання не відповідає рівню підготовки, кількість інформації зменшується. Відповідний доданок має меншу вагу. Через це статистика **infit** може бути нечутлива до екстремальних

відповідей і навіть зовсім їх не «бачити». Чим краще експериментальні дані узгоджуються з моделлю Раша, тим ближче до одиниці значення обох статистик. Прийнятними вважаються значення з проміжку (0.8; 1.2).

Обидві статистики $U_j^{(1)}$ (outfit) та $U_j^{(2)}$ (infit) можна деяким перетворенням (Wilson-Hilferty) трансформувати до стандартного нормального розподілу із середнім 0 та відхиленням 1:

$$t_j^{(k)} = \left(\sqrt[3]{U} - 1\right) \cdot \left(\frac{3}{\sigma(U)}\right) + \frac{\sigma(U)}{3}, \; k = 1, 2, \qquad (4.6)$$

де U – одна із статистик (4.2) або (4.4), а $\sigma(U)$ – відповідне середньо квадратичне відхилення (формули (4.3), (4.5)).

Для статистики (4.6) можна знайти критичні точки. Наприклад, на рівні значущості $\alpha = 0{,}05$ отримуємо $t_{кр} = \Phi^{-1}\left(\frac{1-\alpha}{2}\right) = 1{,}96 \approx 2$. Якщо одна із статистик $t_j^{(1)}$ чи $t_j^{(2)}$ не потрапляє у проміжок $(-2, 2)$, гіпотезу про відповідність профілю відповіді моделі вимірювання відхиляють. Значення $t_j^{(k)} < -2$ вказує на меншу варіативність результатів, ніж передбачає модель. Тобто профіль відповідей аж занадто відповідає ідеальному профілю Гуттмана, де всі нулі знаходяться тільки після одиниць $111\ldots10\ldots0$. При $t_j^{(k)} > 2$ робимо висновок, що профілі відповідей є більш випадковими, ніж за моделлю. Таке трапляється, коли сильні опитані не відповідають на просте питання, або навпаки. Профіль містить багато порушень порядку $01110\ldots10\ldots101$.

Незважена статистика (4.2) може використовуватись для аналізу нахилу емпіричної кривої по відношенню до теоретичної. Можна показати, що коли емпіричні дані знаходяться над теоретичною характеристичною кривою як показано на рис.27 *a*), то значення статистики outfit більше за 1. Емпірична крива пологіша. Таке завдання має насправді меншу дискримінуючу здатність, ніж передбачає модель. Для ситуації на рис.27 *b*) значення статистики відповідно менше за 1. Емпірична крива крутіша. Цей факт також використовують для перевірки припущення моделі Раша про однакову дискримінуючу здатність завдань.

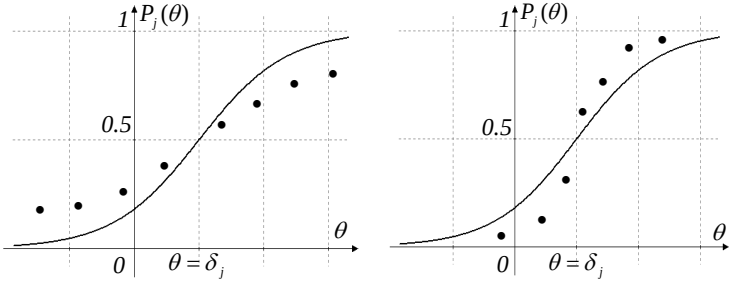


Рис. 27.

Зауважимо, що залишкові статистики не дають інформації про те, як далеко від теоретичної кривої знаходяться емпіричні дані. Наприклад, зважена статистика може мати значення близьке до 1 як при малих, так і при великих відхиленнях окремих значень навколо теоретичної кривої. Про ступінь близькості можна робити висновок з асимптотичної дисперсії середньо квадратичних статистик, яка дорівнює $\frac{2}{n}$. Чим більший об'єм вибірки, тим менше розсіяні значення статистик навколо 1.

Також ці статистики не дають значення коефіцієнта дискримінуючої здатності, а лише вказують на можливі відхилення від теоретичного. Більше інформації про нього можна отримати, аналізуючи коефіцієнти кореляції між загальним балом та відповіддю на питання у рамках класичної теорії. Тому більшість програмних продуктів забезпечують такий аналіз.

Наведемо ще кілька статистик, які використовуються для перевірки відповідності даних побудованій моделі. Це, у першу чергу, ***Lz***-статистика, досліджена Reise (1990):

$$l_z = \frac{l(\theta) - M\left[l(\theta)\right]}{\sqrt{D\left[l(\theta)\right]}}, \tag{4.7}$$

де $l(\theta) = \sum_{i=1}^{n}\left[x_{ij} \ln P_{ij} + (1 - x_{ij}) \ln(1 - P_{ij})\right],$

$$M\left[l(\theta)\right] = \sum_{i=1}^{n}\left[P_{ij} \ln P_{ij} + (1 - P_{ij}) \ln(1 - P_{ij})\right],$$

$$D\left[l(\theta)\right]=\sum_{i=1}^{n}P_{ij}(1-P_{ij})\left[\ln\left(\frac{P_{ij}}{1-P_{ij}}\right)\right]^2.$$

Статистика ***Lz*** має близький до стандартного нормального розподіл, якщо отримані задовільні оцінки. Від'ємне значення ***Lz*** вказує на неймовірні профілі, а додатне – на профілі більш послідовні, ніж очікує модель.

Tatsuoka та Linn (1983) запропонували шість статистик застереження для виявлення відхилень у відповідях, дві з яких ***Eci2z*** та ***Eci4z***, засновані на порівнянні коваріацій, успішно використовуються в IRT:

$$Eci2z=\frac{\sum_{i=1}^{n}(P_{ij}-x_{ij})(G_i-\mu_G)}{\sqrt{\sum_{i=1}^{n}P_{ij}(1-P_{ij})(G_i-\mu_G)^2}}, \tag{4.8}$$

$$Eci4z=\frac{\sum_{i=1}^{n}(P_{ij}-x_{ij})(P_{ij}-\mu_P)}{\sqrt{\sum_{i=1}^{n}P_{ij}(1-P_{ij})(G_i-\mu_P)^2}}. \tag{4.9}$$

Тут $G_i=\frac{1}{k}\sum_{i=1}^{n}P_{ij}$, $\mu_G=\frac{1}{n}\sum_{i=1}^{n}G_i$, $\mu_P=\frac{1}{n}\sum_{i=1}^{n}P_{ij}$. Статистика ***Eci2z*** порівнює профілі відповідей з середньою ймовірністю відповідей на тестові завдання, в той час як ***Eci4z*** порівнює профілі з очікуваною ймовірністю в залежності від моделі Раша. Малі значення цих статистик, близькі до 0, вказують на добру відповідність даних з пропонованою моделлю.

4.2. Перевірка гіпотез

Якщо у попередніх статистиках фігурували окремі первинні бали, то для застосування критерію χ^2 для перевірки відповідності даних обраній моделі потрібно згрупувати учасників тестування по шкалі θ за діапазонами рівня підготовленості. Нехай учасники тестування діляться на T груп так, що всередині кожної групи учасники мають однаковий рівень підготовленості θ_t. Всього усере-

дині групи з номером t міститься m_t учасників ($t = 1,2,\ldots,T$). В межах кожної групи r_{tj} учасників відповідають вірно на j–те завдання тесту. Для групи з рівнем підготовленості θ_t емпірична ймовірність вірної відповіді на j–те завдання тесту дорівнює $p_{tj} = \dfrac{r_{tj}}{m_t}$, а теоретична ймовірність $P(\theta_{tj})$ обчислюється відповідно до обраної моделі. Yen (1981) запропонувала для дослідження використовувати статистику Q_{1j}, яка базується на квадратах різниць і є по суті статистикою хі-квадрат Пірсона:

$$Q_{1j} = \sum_{t=1}^{T} m_t \cdot \frac{\left[p_{tj} - P(\theta_{tj})\right]^2}{P(\theta_{tj})(1 - P(\theta_{tj}))}. \qquad (4.10)$$

Величина (4.10) має розподіл χ^2 з $(T - b)$ ступенями волі, де b – кількість параметрів моделі, наприклад для моделі Раша $b = 1$.

Щоб дослідити основну гіпотезу H_0 про відповідність отриманих емпіричних даних побудованій характеристичній кривій, знаходять спостережене $Q_{1j}^{спос}$ та критичне $Q_{1j}^{крит}$ (за таблицями) значення критерію. Задають рівень надійності α. За результатами порівняння приймається одне з двох рішень:

1) якщо $Q_{1j}^{спос} > Q_{1j}^{крит}$, то нульова гіпотеза на даному рівні значущості відкидається, тобто модель не узгоджується з експериментальними даними;

2) якщо $Q_{1j}^{спос} \le Q_{1j}^{крит}$, то нема підстав для відхилення нульової гіпотези, тобто модель добре описує дані.

Інша статистика, заснована на відношенні функцій вірогідності, була запропонована McKinley та Mills (1985):

$$G_j^2 = -2\ln\frac{L_{j1}}{L_{j0}} = 2\sum_{t=1}^{T}\left[r_{tj}\ln\left(\frac{p_{tj}}{P(\theta_{tj})}\right) + (m_t - r_{tj})\ln\left(\frac{1-p_{tj}}{1-P(\theta_{tj})}\right)\right]. \qquad (4.11)$$

Ця статистика теж має наближено розподіл χ^2. Проблемою статистик Q_1 та G^2 є те, що групування учасників відбувається на основі оцінок θ. Є статистики, що базуються на первинних балах.

Статистика G^2 дає більшу помилку I роду.

Зробити висновок про відповідність завдань та усього тесту вибраній теорії також можна на основі статистики

$$\chi^2 = \sum_{j=1}^{k} \chi^2(j), \quad \text{де} \quad \chi^2(j) = \sum_{t=0}^{k} m_t \frac{\left[p_{tj} - P(\theta_{tj})\right]^2}{P(\theta_{tj})}. \qquad (4.12)$$

Тут кількість груп дорівнює кількості усіх можливих балів. Величина $\chi^2(j)$ має розподіл хі-квадрат з числом ступенів волі k, коли є максимальна кількість груп $(k+1)$. Якщо малочисельні групи (кількість учасників менше 5) об'єднати, число степенів свободи зменшиться. Дана статистика χ^2 має максимальне число ступенів свободи $k(k+1)$.

5. ПРОГРАМНІ ЗАСОБИ ДЛЯ АНАЛІЗУ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ

5.1. Можливості пакету програм ІТАР

Для оцінки параметрів обраних моделей на основі даних тестування для певної вибірки використовують різні статистичні процедури, що вимагають великих об'ємів розрахунків. З кінця 70-х років минулого століття з'явилось багато комп'ютерних програм, у яких ці процедури було реалізовано.

Добре відомою у свій час для обчислень за методом максимальної вірогідності на основі трипараметричної моделі була програма LOGIST (перша версія 1976 р.). Вона дозволяє одночасно оцінювати параметри для всіх завдань та отримувати оцінки латентної характеристики для всіх учасників тестування за сумісною процедурою максимальної вірогідності JML (*joint maximum likelihood procedure*). Недоліком такої процедури є відкрите питання про слушність отриманих оцінок параметрів завдань та необхідність мати великі вибірки (більше 1000 опитаних) для забезпечення більш точних оцінок. У програмі BILOG (1984 р.) для оцінки параметрів за трипараметричною моделлю використовується процедура маргінальної максимальної вірогідності MML (*marginal*

maximum likelihood procedure). Оцінки, отримані за цією процедурою, є слушними. Програма BILOG та її вдосконалені версії широко використовується у країнах Північної Америки та Європи. Для однопараметричної моделі (або ж моделі Раша) початківцям рекомендували використовувати програму BICAL (1979 р.), оскільки її результати дуже прості для інтерпретації. Розробники BICAL запропонували до процедури JML ввести додатковий множник, який дозволяє отримати майже слушні оцінки параметрів. Умовну процедуру максимальної вірогідності CML (*conditional maximum likelihood procedure*) для моделі Раша використали розробники програм RUMM (1990 р.) та WINMIRA (2001 р.). До переліку популярних сьогодні програм можна також віднести MULTILOG (1991 р.), PARSCALE (1997 р.), продукти компанії ASC (Assessment System Corporation) та багато інших. Користувачеві залишається лише вибрати потрібну модель, програмний засіб для її реалізації та правильно проінтерпретувати отримані результати обчислень.

Далі розглянемо деякі можливості пакету ITAP (Item and Test Analysis Package) компанії ASC. Пакет містить шість програм, розроблених у 1995-1998 рр. та вдосконалених для Windows, які дозволяють здійснювати всебічний аналіз результатів тестування, анкетування, соціологічних чи психологічних опитувань у рамках класичної та сучасної тестової теорії.

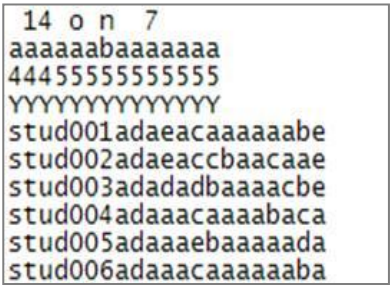


Рис. 28.

Дані для аналізу у трьох програмах (ITEMAN, RASCAL та XCALIBRE) створюються як текстові ASCII файли і мають однаковий формат з деякими відмінностями, що залежать від типу тес-

ту. На рис.28 наведено фрагмент вхідного файлу для дихотомічних завдань. У першому контрольному рядку через пропуск потрібно зазначити кількість завдань (до 750), символ для позначення пропущених відповідей (o), символ для позначення питань, на які опитуваний не встиг відповісти (n), кількість символів для ідентифікації опитаних (до 80). Другий рядок – ключ правильних відповідей на кожне завдання, третій – кількість альтернатив у кожному завданні (від 2 до 9). У четвертому рядку літерою Y позначають завдання, які включаються до аналізу, а N – ті, що не включаються. Далі наступним суцільним рядком – відповіді кожного опитаного, причому кількість опитаних для всіх програм необмежена.

Програма ITEMAN забезпечує традиційний аналіз тесту або опитування у рамках класичної теорії. Вона дозволяє аналізувати дані, отримані у дихотомічній та у багатопозиційних шкалах (типу шкали Лайкерта). Передбачена можливість аналізу питань з кіль-

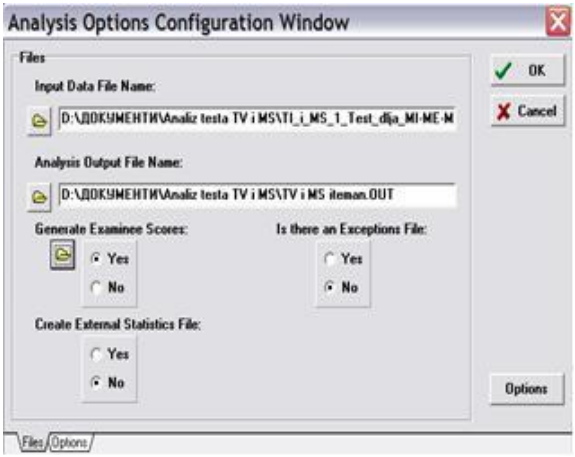


Рис. 29.

кома правильними відповідями. Для цього у вікні програми ***Files*** (рис.29), у якому визначаються всі робочі файли, потрібно вибрати кнопку *Yes* у полі *Exceptions Files* та сформувати додатковий файл для таких питань.

У всіх програмах передбачена можливість формування вихідного файлу у вигляді, зручному для експорту у інші статистичні пакети для подальшої обробки (*External Statistics File*). Параметри

аналізу задаються у вікні ***Options***. Тут є можливість вибрати вид кореляційних зв'язків, поділ опитуваних на слабку та сильну групи, аналіз тесту у підгрупах, уточнення параметрів завдань для коротких тестів. В ITEMAN, на відміну від інших програм пакету, не розрізняються пропущені питання та ті питання, на які опитуваний не встиг відповісти. Для аналізу питань, відповідь на які опитуваний пропустив, є три різні процедури. Вихідний текстовий файл містить загальні статистичні показники тесту, кожного завдання та кожної альтернативи залежно від обраної опції аналізу. Аналіз характеристик кожного завдання дозволяє виявити ті з них, які не коректно працюють у тесті і потребують вдосконалення.

Наприклад, завдання 13 (рис.30) має від'ємну кореляцію з усім тестом та низький показник дискримінуючої здатності. Слабкі студенти відповідали на це питання краще, ніж сильні (можливо вгадували). Потрібно перевірити, чи не було помилки при введенні ключа, оскільки більшість опитаних обирала відповідь *c)* замість правильної *a)*. Якщо незадовільні характеристики завдання не пояснюються механічними помилками, то таке завдання з тесту потрібно вилучити.

		Item Statistics			Alternative Statistics					
Seq. No.	Scale -Item	Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	Endorsing High	Point Biser.	Key
13	0-13	.15	.07	-.17	A	.15	.11	.18	-.17	*
					B	.33	.50	.18	-.49	
	CHECK THE KEY				C	.25	.17	.41	.11	?
	a was specified, c works better				D	.21	.17	.18	-.18	
					E	.06	.06	.06	-.13	
					Other	.00	.00	.00		
14	0-14	.54	.49	.29	A	.54	.28	.76	.29	*
					B	.00	.00	.00		
					C	.02	.00	.06	-.01	
					D	.06	.00	.06	-.09	
					E	.38	.72	.12	-.68	
					Other	.00	.00	.00		

Рис. 30.

Програма RASCAL дозволяє аналізувати тест у рамках моделі Раша. Тут за ідеологією моделі Раша досліджується відповідність емпіричних даних моделі, чого немає у інших програмах даного пакету. Для оцінки латентних параметрів рівня складності

завдань та рівня підготовленості опитаних у RASCAL використовується метод JML з деякими уточненнями для кращого узгодження з трипараметричною моделлю. Програма працює лише з дихотомічними даними, всі пропущені питання опрацьовуються як питання з неправильною відповіддю. Програма має два головних вікна ***Files*** та ***Options***, у яких формуються усі необхідні робочі файли та визначаються опції аналізу. Є можливість обирати початок відліку на шкалі Раша: за середнім рівнем підготовки або за середньою складністю завдань. Функціональну залежність можна вибрати у вигляді логістичної або нормальної кривої. Можна задати максимальну кількість ітерацій, внести поправку для невеликих вибірок, задати довільне лінійне перетворення шкали логітів. Вихідний текстовий файл, крім оцінок параметрів тесту (рис.31), містить графічне зображення шкали з відміченими значеннями латентних параметрів підготовленості та складності завдань (рис.32), графіки характеристичної та інформаційної функцій тесту. Оцінки рівня підготовленості кожного опитаного можна вивести у окремому файлі, який потрібно задати у полі *Examinee Scores* вікна ***Files***.

Sorted in Item Difficulty Order

Item	Difficulty	Std. Error	Chi Sq.	df	Scaled Diff
3	-3.331	0.906	2.683	7	70
1	-2.240	0.569	2.610	7	80
5	-1.694	0.467	3.318	7	85
11	-0.299	0.331	6.014	7	97
8	-0.100	0.323	5.115	7	99
9	-0.004	0.320	9.331	7	100
6	0.275	0.314	4.498	7	103
14	0.366	0.313	3.606	7	103
10	0.456	0.312	8.353	7	104
4	0.456	0.312	7.564	7	104
12	1.009	0.319	4.074	7	109
2	1.205	0.325	2.740	7	111
7	1.412	0.334	4.657	7	113
13	2.487	0.418	9.649	7	123

Рис. 31.

На рис.31 крім оцінок складності завдань (другий стовпчик) є значення статистики χ^2 (четвертий стовпчик), яка вказує на відповідність емпіричних даних побудованій теоретичній моделі. Тут на рівні значущості 0,05 для 7 ступенів свободи можна стверджувати, що завдання не відповідатиме моделі, якщо значення χ^2

перевищуватиме 14,1. Як бачимо, жодне із завдань не потрапило у критичну область. За шкалою логітів на рис.32 можна визначити співвідношення між рівнем підготовленості опитаних та складністю запропонованих завдань. Можна зробити висновок, що тест майже збалансований, складність більшості завдань відповідає рівню підготовленості. Три найлегші завдання з тесту можна вилучити, оскільки з ними впораються навіть найгірше підготовлені опитані.

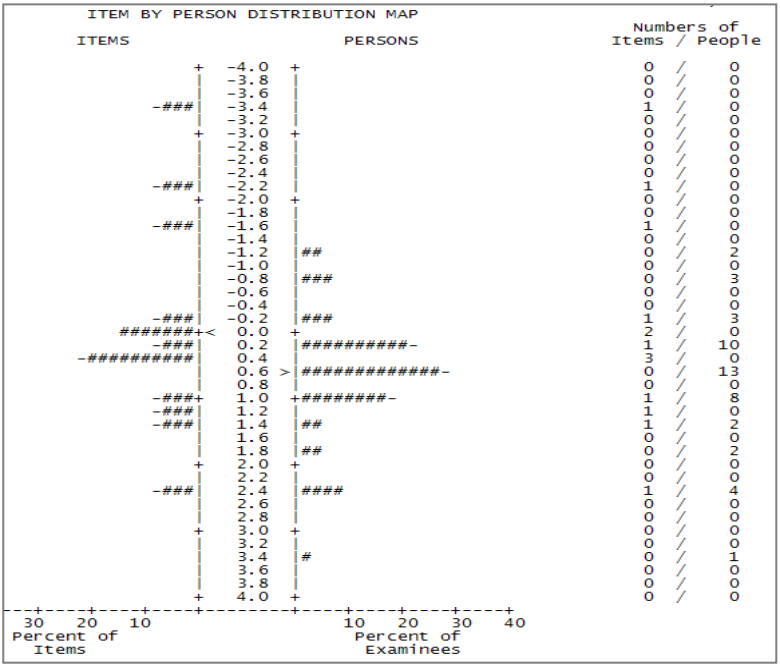


Рис. 32.

Недоліком усіх програм пакету ITAP можна вважати незручну графічну інтерпретацію отриманих результатів. Не передбачена можливість побудови характеристичних та інформаційних функцій кожного завдання. Для їх побудови потрібно статистичні показники тесту зберегти у окремому файлі, створеному у полі *External Statistics* вікна ***Files***. Експортуючи ці дані, наприклад, у Excel, можна побудувати характеристичні криві кожного завдання.

Проаналізувати результати тесту за дво- та трипараметрич-

ною моделями IRT з урахуванням дискримінуючої здатності кожного завдання та угадування можна за допомогою програми XCALIBRE. Вона використовує маргінальний метод максимальної вірогідності (MML) для оцінки латентних параметрів обох моделей, що робить можливим її застосування для аналізу коротких тестів (менше 25 питань) та невеликих масивів опитаних (менше 1000 осіб). Програма працює лише з дихотомічними даними. Для аналізу тут суттєво, чи тестований пропустив завдання, чи не встиг відповісти. Призначення вікон ***Files*** та ***Options*** тут таке ж, як у попередніх програмах. При визначенні опцій аналізу потрібно вибрати тип моделі, початкові розподіли для оцінок параметрів та кількість ітерацій. Якщо генерується файл з оцінками опитаних, то при запуску опції *Analyze* потрібно вибрати метод побудови оцінки латентного параметра рівня підготовленості (за замовчуванням це *Maximum-Likelihood*).

FINAL ITEM PARAMETER ESTIMATES

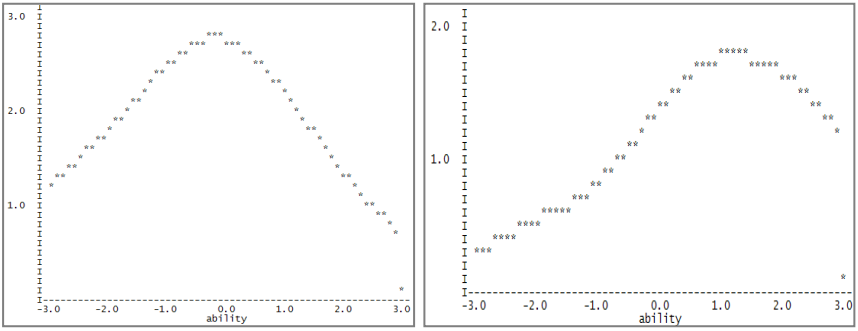
Item	Lnk	Flg	a	b	c	Resid	PC	PBs	PBt	N
1			0.72	-2.41	0.26	0.35	0.94	0.19	0.35	48
2			0.74	1.86	0.25	0.45	0.36	0.45	0.25	48
3	-- Deleted --									
4			0.71	0.71	0.26	0.13	0.53	0.57	0.42	48
5			0.72	-1.90	0.26	0.13	0.90	0.36	0.45	48
6			0.69	1.05	0.32	0.61	0.57	0.33	0.12	48
7			0.75	1.78	0.22	0.39	0.31	0.55	0.43	48
8			0.72	-0.13	0.25	0.16	0.65	0.30	0.56	48
9			0.72	0.00	0.25	0.20	0.64	0.45	0.53	48
10			0.72	0.54	0.24	0.16	0.53	0.47	0.56	48
11			0.66	-0.14	0.28	0.49	0.69	0.15	0.29	48
12			0.71	2.41	0.30	0.38	0.40	0.14	-0.02	48
13			0.75	3.00	0.19	0.64	0.15	-0.00	-0.16	48
14			0.72	0.70	0.27	0.34	0.54	0.52	0.39	48

Рис. 33.

Вихідний текстовий файл містить оцінки параметрів складності (***b***), дискримінуючої здатності (***a***) та угадування (***c***) для кожного завдання (рис.33), їх похибки, аналіз запитань та альтернатив, графіки характеристичної та інформаційної функцій тесту. Якщо обрано двопараметричну модель, то параметр $c = 0$.

З вихідного файлу для трипараметричної моделі (рис.33) бачимо, що завдання 3 було вилучено з аналізу як не інформативне,

оскільки майже всі опитані дали вірну відповідь, крім однієї пропущеної. У стовпчику *Flg* для проблемних питань можлива поява міток: *P* – якщо значення оцінюваних параметрів виходять за межі моделі ($a < 0{,}3$, $b < -2{,}95$ або $b > 2{,}95$, $c > 0{,}4$); *K* – якщо деяка з альтернатив має кращу кореляцію з тестом, ніж правильна відповідь (невірний ключ); *R* – якщо статистика відповідності моделі (стовпчик *Resid*) перевищує 2,0. У стовпчику *PC* маємо відсоток тих, хто правильно відповів на всі питання серед тих, хто дав вірну відповідь на дане питання. Наступні стовпчики містять інформацію про кореляцію відповідей на кожне завдання з відповідями за весь тест (*PBs*) та з отриманими оцінками латентних параметрів (*PBt*).



a *b*

Рис. 34.

На рис.34 зображені інформаційні функції тесту у двовимірній (*a*) та тривимірній (*b*) моделях. Аналізуючи їх, приходимо до висновку, що врахування параметра угадування значно зменшує роздільну здатність тесту та його інформативність. Для слабкої групи опитаних результати, швидше за все, досягнуто за рахунок угадування. Тест найбільш інформативний саме для сильної групи, рівень підготовленості якої дорівнює або більше 1 логіта.

Наступні три програми пакету ITAP на основі попередньо знайдених статистичних характеристик тесту дозволяють оцінити його надійність (TESTINFO), валідність (TESTVAL) та порівняти оцінки рівня підготовленості, отримані за різними методиками (SCOREALL).

5.2. Обробка результатів тестування у WINSTEPS

Популярною серед користувачів є також програма WINSTEPS (Windows версія BIGSTEPS, 1991 р.), яка дозволяє для дослідження тестів з дихотомічними чи політомічними завданнями використовувати моделі:

- однопараметричну (Georg Rasch);
- rating scale model (Andrich);
- partial credit model (Masters);
- paired comparison model (Bradley-Terry);
- success model (Glas);
- failure model (Linacre),

а також різні комбінації цих моделей.

Для оцінки параметрів тут також використовується процедура JML з початковим наближенням, яке отримується за процедурою PROX. Популярності цій програмі додає те, що безкоштовно поширюється її академічна версія MINISTEP, у якій можна аналізувати до 25 завдань на вибірці до 75 опитаних.

Програма працює з даними, які можна підготувати у різних статистичних пакетах. Для прикладу розглянемо матрицю відповідей, підготовлену в Excel. Стовпчики відповідають завданням, рядки – учасникам тестування. У кожній клітинці – кількість набраних балів (0 або 1 для дихотомічних завдань, від 0 до максимальної кількості балів за завдання – для політомічних).

Рис. 35.

Конвертування даних у програму WINSTEPS здійснюється наступним чином: після запуску WINSTEPS необхідно на питання *Would you like help setting up your analysis?* відповісти *No*. Потім вибрати пункт *Excel/RSSST*, після чого з'явиться вікно *Select data*

to be converted to Winsteps format, в якому необхідно вибрати кнопку Excel (рис.35). Після чого з'явиться вікно *Excel Input for Winsteps*, де обираємо кнопку *Select Excel file* (рис.36).

Рис. 36.

Далі у списку файлів знаходимо необхідний файл для конвертування та відкриваємо його. В результаті чого відкривається вікно *Excel Input for Winsteps* (рис.37), де потрібно правильно розташувати завдання та випробуваних.

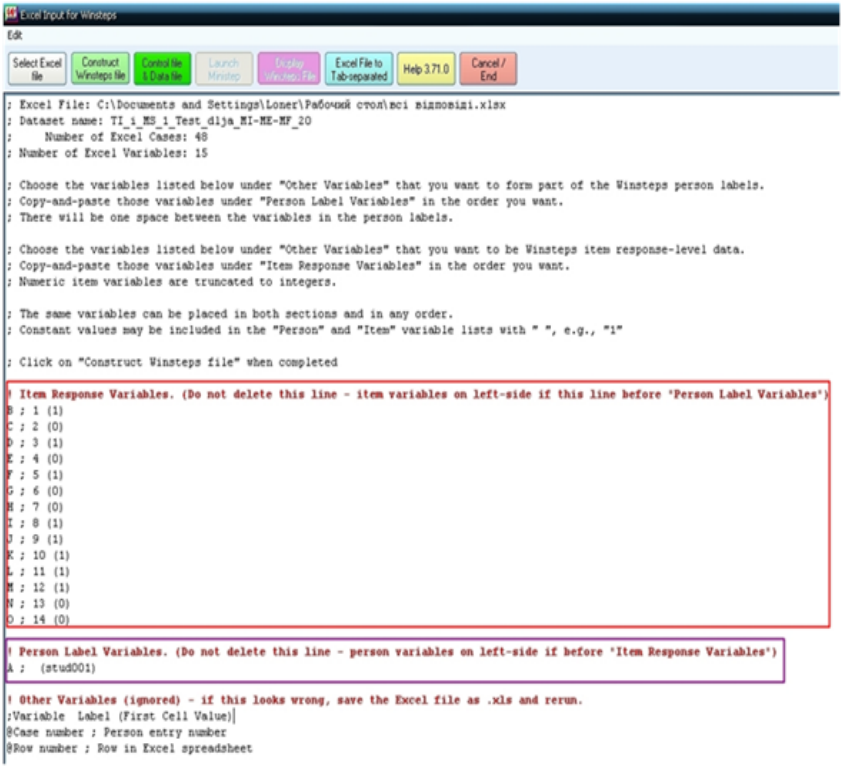


Рис. 37.

Необхідно скопіювати імена змінних, що відповідають номерам завдань, які розташовані нижче стрічки *;Variable Format Label* (в кінці файлу), та вставити їх після стрічки *!Item Response Variables*, змінну *A ; (stud001)*, яка відповідає опитаним, потрібно вставити після стрічки *!Person Label Variables*. В результаті чого відредагований файл має вигляд як на рис.37.

Потім натискаємо кнопку *Construct Winsteps file*, надаємо ім'я переконвертованому в текстовий формат файлу, та зберігаємо його. В результаті цих дій, на екран виводиться збережений файл (рис.38).

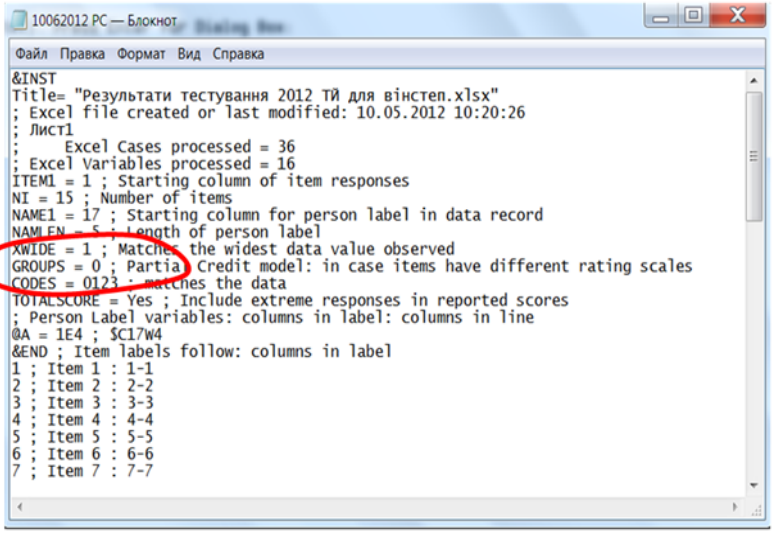


Рис. 38.

Командні рядки формуються автоматично. Якщо матриця відповідей була дихотомічною, за замовчуванням відбуватиметься аналіз відповідно до моделі Раша. Для завдань політомічного типу можна використати моделі RSM (якщо всі завдання мають однакову кількість категорій) та PCM. У WINSTEPS обидві моделі належать до однієї групи, що визначається командою MODELs=R (або її відсутністю). Для аналізу у PCM потрібно у командному рядку вказати параметр GROUPS=0. Його відсутність або параметр GROUPS=" " забезпечує аналіз у моделі RSM. Будь-який текст у рядку після крапки з комою сприймається як довідкова інформація.

Після конвертування даних переходимо до їх аналізу. У цьому ж вікні *Excel Input for Winsteps* натискаємо кнопку *Launch Winsteps* (або *Ministep*). Після її натиснення програма повертає нас до головного вікна, де для аналізу вибраний збережений раніше файл (рис.39).

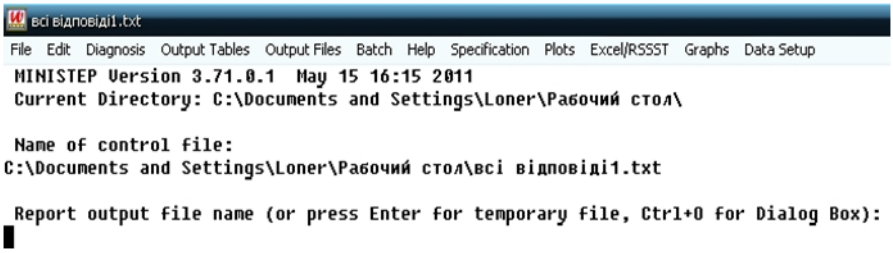


Рис. 39.

Щоб розпочати аналіз потрібно двічі натиснути кнопку *Enter*, програма виведе на екран загальні дані про кількість завдань та випробуваних, середні значення оцінених параметрів та дисперсію. Для більш докладнішого аналізу завдань та учасників потрібно скористатися пунктом головного меню *Output Tables* з різними табличними звітами або пунктом *Graphs* (рис.40) з графічними звітами.

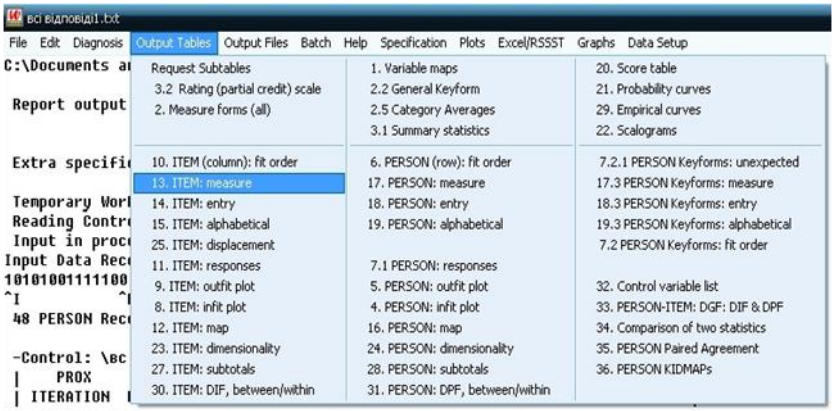


Рис. 40.

Результати конвертування вхідних даних у вимірювання Раша для всіх завдань тесту знаходяться у звіті ITEM STATISTICS

(меню *Output Table, 13. Item: measure*) (рис.41).

ITEM STATISTICS: MEASURE ORDER

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

Рис. 41.

У першому стовпчику (ENTRY NUMBER) вказані номери завдань, у другому (TOTAL SCORE) – кількість правильних відповідей даних на це питання серед 48 випробуваних, у третьому (TOTAL COUNT) – загальна кількість усіх відповідей. Результати вимірювання складності завдань у логітах (стовпчик MEASURE) наведені у порядку спадання (наприклад, завдання 13 майже на 1,17 логіти складніше, ніж завдання 7 і т.д.). У стовпчику MODEL S.E. наведена похибка вимірювання на основі моделі Раша, а у рядках MEAN та S.D. – середні значення та стандартні відхилення для значень у відповідних стовпчиках.

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

менші за 0,5 теж небажані, бо свідчать про «інформаційну перевантаженість» питання. Аналіз починають із питань з високим значенням MNSQ. У полі ZSTD наводяться стандартизовані значення MNSQ. Прийнятними є значення від –2 до +2.

У стовпчику PT-MEASURE CORR. наводяться значення коефіцієнта кореляції. Даний коефіцієнт може приймати значення від –1 до +1. Він розглядається як деякий показник надійності та валідності і може бути використаний для визначення, доопрацювання а можливо і виключення слабко узгоджених завдань. Наприклад, завдання 13 має від'ємну кореляцію з усім тестом. Слабкі студенти відповідали на це питання краще, ніж сильні (можливо вгадували).

Можна також отримати звіт з характеристиками учасників тестування PERSON STATISTICS (меню *Output Tables, 17. Person: measure*). На рис.42 його фрагмент.

PERSON STATISTICS: MEASURE ORDER

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

Рис. 42.

При всіх задовільних значеннях більшості показників насторожує значення 9,9 показника MNSQ OUTFIT для учасника під номером 27. Виявляється, що це саме той студент, якому вдалося за 5 хвилин дати кілька правильних відповідей, що майже напевне вказує на угадування.

У WINSTEPS можна отримати різні графічні звіти: характеристичні криві, інформаційні функції тощо. На рис.43 побудовані характеристичні криві усіх завдань деякого дихотомічного тесту, аналіз взаємного розміщення яких допомагає вдосконалити тест як систему завдань зростаючої складності.

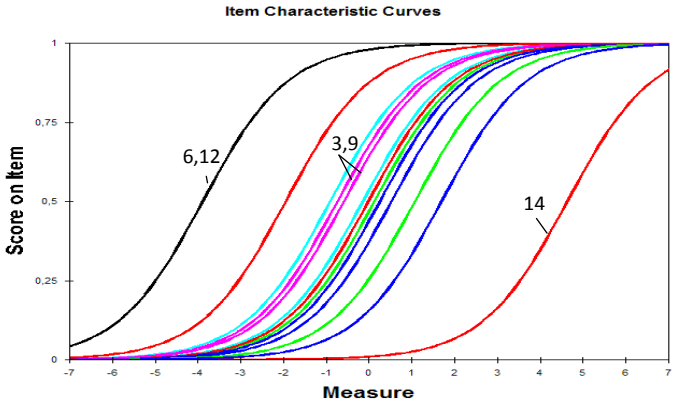
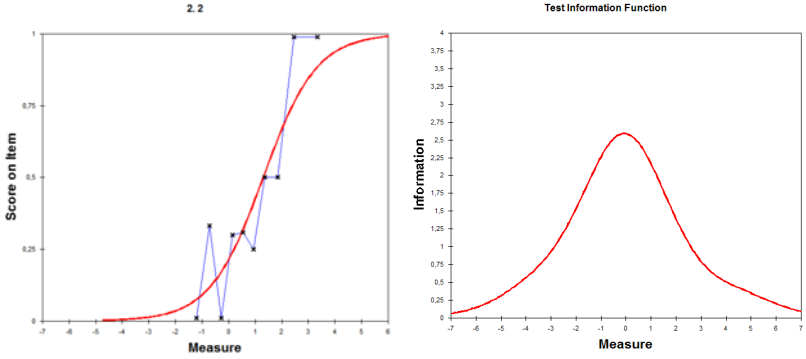


Рис.43.

У даному тесті більшість завдань зосереджені в області середньої та меншої за середню складності. Потрібно додати складніші завдання, щоб рівномірно заповнити область від 0 до 5 логітів. Характеристичні криві для завдань 6 та 12 накладаються, тому одне з них можна вилучити. Такі завдання можуть бути використані для паралельних тестів. Для ідеального тесту характеристичні криві повинні рівномірно заповнювати весь інтервал (–5;+5).



a) *b)*

Рис. 44.

По кожному завданню та по тесту загалом можна отримати графічне представлення відповідності даних обраній моделі (рис. 44, *a*). Інформаційні функції завдань та тесту виглядають як на рис.44, *b*). Аналогічні звіти отримуємо для політомічних завдань.

Підготовку вхідних даних безпосередньо у WINSTEPS здійснюють за допомогою пункту меню *Data Setup*, де забезпечено зручне автоматичне формування командних рядків робочого файлу (рис.45). Крім того, за допомогою пункту меню *Plots* можна отримати табличне та графічне представлення результатів в Excel. Використовуючи пункт *Output Files*, можна зберегти вихідні файли в SPSS або в R-Statistics.

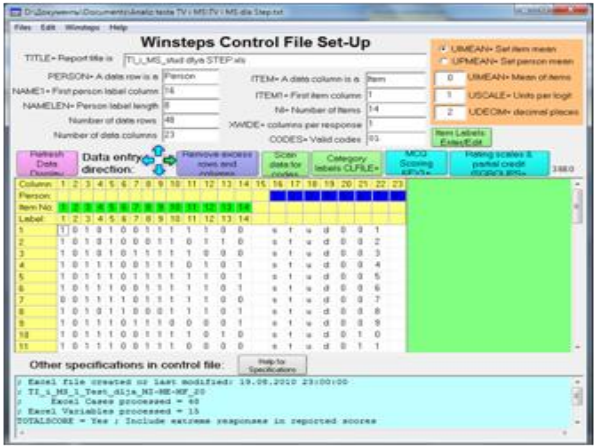


Рис. 45.

У посібнику не розглядались методи класичної теорії тестів. Вимірювання за класичною схемою та за допомогою моделей сучасної теорії тестів мають різну внутрішню логіку та різні обчислювальні процедури. Саме тому для посилення надійності педагогічних вимірювань варто не замінювати, а доповнювати ці два підходи один одним.

ЛІТЕРАТУРА

1. Вимірювання в освіті: Підручник / За редакцією О.В. Авраменко.– Кіровоград: Лисенко В.Ф., 2011. – 360 с.

2. Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. – М.: Прогресс, 1976. – 494 с.

3. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1998. – 400 с.

4. Дубина И.Н. Математические основы эмпирических социально–экономических исследований. – Барнаул: Изд-во Алт.ун-та, 2006. – 263 с.
5. Звонников В.И. Современные средства оценивания результатов обучения: учеб. Пособие для студ. высш. учеб. заведений / В.И. Звонников, М.Б. Челышкова. – М.: Издательский центр «Академия», 2007. – 224 с.
6. Ким В.С. Тестирование учебных достижений. Монография. – Уссурийск: Изд-во УГПИ, 2007. – 214 с.
7. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов. – М.: Логос, 2010. – 668 с.
8. Майоров А.Н. Теория и практика создания тестов для системы образования. – М.: Интеллект – Центр, 2002. – 296 с.
9. Наследов А. Д. Математические методы психологического исследования. Анализ и интерпретация данных. Учеб. пособие. – СПб.: Речь, 2004. – 392 с.
10. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. – М.: Прометей, 2000. – 168 с.
11. Новиков Д.А. Статистические методы в педагогических исследованиях (типовые случаи). – М.: МЗ–Пресс, 2004. – 67 с.
12. Челышкова М.Б. Теория и практика конструирования педагогических тестов. – М: Логос, 2002. – 432 с.
13. Baker F.B. The Basics of Item Response Theory. – Portsmouth NH: Heinemann Educational Books, 1985. – 131 pp.
14. Brennan R. Educational Measurement. – Westport, CT: Praeger, 2006. – 796 pp.
15. De Mars Ch. Item response theory. – Oxford University Press, 2010. – 131 pp.
16. Hambleton R.K., Swaminathan H., Rogers H. J. Fundamentals of Item Response Theory. – Newbury Park, CA: Sage, 1991. –175 pp.
17. Ostini R., Nering M.L. Polytomous item response theory models. – Australia: Measured Progress, 2006. – 120 pp.
18. Reckase M.D. Multidimensional Item Response Theory. – New York: Springer, 2009. – 354 pp.
19. Wright B.D., Stone M.H. Best Test Design. – Chicago, MESA PRESS, 1979. – 222 pp.
20. www.winsteps.com.
21. www.assess.com.