9. ВСТУП ДО ТЕОРІЇ IRT

Англійську назву теорії *Item Response Theory* (IRT) не можна перекласти українською дослівно. Пропонувалися різні варіанти українського відповідника цій назві, однак жоден з них не є цілком прийнятним. Зокрема, це стосується терміну «сучасна теорія тестування», який виник, мабуть, на противагу терміну «класична теорія тестування» (Classical Test Theory, CTT) яким позначають теорію, що ґрунтується на класичній моделі тестової оцінки, яку ми розглянули у главі 5, і за межі якої досі не виходили. Але термін «сучасний» зовсім не є таким, який можна протиставити терміну «класичний». IRT – це не теорія, покликана загалом замінити класичну теорію. У порівнянні з останньою вона має не тільки явні переваги, але й істотні недоліки, передусім в плані практичного застосування. Тому зараз обидві теорії успішно співіснують у психометрії, зокрема, в освітніх вимірюваннях.

Розглянуті нами раніше характеристики тестових завдань не є достатньо повними. Наприклад, вони не несуть у собі інформацію про те, як розподіляються відповіді на завдання для осіб з визначеним рівнем вимірюваної якості.

IRT заснована на математичних моделях, які здатні показати, як особи з різними рівнями вираженості вимірюваної якості повинні відповідати на завдання тесту.

У цій книзі ми лише познайомимося з фундаментальними основами даної теорії.

Функція відповіді на завдання. Надалі замість слів «вимірювана якість», які означають рису або конструкт, будемо використовувати термін «латентна характеристика», деталізуючи цей термін по мірі викладу матеріалу. Також надалі ми вважатимемо, що завдання тесту, яке аналізується, є дихотомічним.

Ми справедливо вважаємо, що загалом на певне завдання відповідають правильно особи з високим рівнем латентної характеристики, і відповідають неправильно особи з низьким рівнем латентної характеристики. Якщо відкласти вздовж горизонтальної осі рівні латентної характеристики, а вздовж вертикальної осі – ймові-

рність правильної відповіді, то в ідеальному випадку для даного завдання існує така точка – значення θ' рівня латентної характеристики θ, що для осіб з рівнем, нижчим ніж θ', ймовірність правильної відповіді дорівнює нулю, тобто всі особи з нижчим від θ'рівнем відповідають на завдання неправильно. І навпаки, для усіх тих, чий рівень латентної характеристики вищий від θ', відповідають на завдання правильно. і ймовірність відповіді особи з таким рівнем дорівнює одиниці. Ця ситуація зображена на малюнку 9.1.

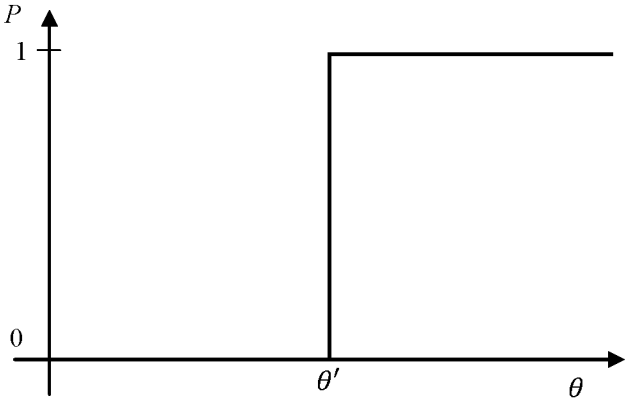

Рис. 9.1. Східчаста функція ICC

Крива, яка відображає залежність ймовірності правильної відповіді на завдання від рівня латентної характеристики особи, називається *функцією відповідей на завдання* (Item Response Function, IRF). У випадку, коли завдання є дихотомічним, ця крива збігається з функцією, яку називають *характеристичною кривою завдання* (Item Characteristic Curve, ICC). Форма кривої, зображеної на рисунку 9.1, майже ніколи не зустрічається на практиці. Зазвичай серед осіб з рівнем нижчим від θ' знаходяться ті, які відповіли на завдання правильно, і, навпаки, серед тих. хто має рівень вищий, ніж θ', знайдуться особи, які відповіли на завдання неправильно. Особливо це стосується тестів рівня навчальних досягнень. Більш того, розумно припустити, що подібні відхилення зустрічаються частіше у осіб з рівнем латентної характеристики,

близьким до θ', а для осіб з екстремально низьким чи високим рівнем подібне спостерігається рідше. Цій ситуації відповідає S-подібна крива, зображена на рисунку 9.2. Криві з такою формою називають *логістичними*. Вони часто слугують моделями для опису процесів у різних галузях. Зокрема, таку форму має крива функції нормального розподілу.

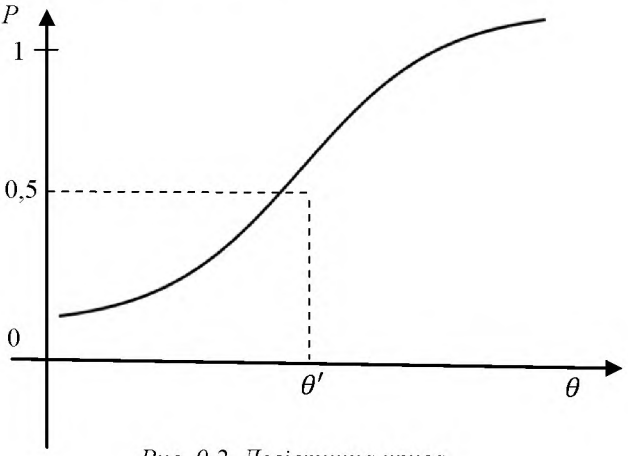

Рис. 9.2. Логістична крива

Але де тепер на цій кривій знаходиться точка, яка відповідає значенню θ', зміст якого можна трактувати так само, як для східчастої кривої з рисунку 9.1? Очевидно, це точка перегину кривої, і їй відповідає ймовірність правильної відповіді 0,5.

Важливо правильно інтерпретувати ймовірність правильних відповідей. Виділимо з популяції субпопуляцію тих осіб, які мають однаковий рівень латентної характеристики, наприклад, $\theta = 2$. Так підпопуляцію назвемо *гомогенною*. Нехай характеристична функція відповідей показує, що цьому рівню відповідає ймовірність 0,87. Це означає, що ймовірність правильно відповісти на завдання для особи з вказаної гомогенної субпопуляції дорівнює 0,87.

Одномірність і локальна незалежність. В IRT постулюється дві важливі концепції – локальної незалежності та одномірності.

Незалежність двох подій трактується як той факт, що те, що одна подія відбулася, ніяк не впливає на ймовірність відбутися для іншої події. Для незалежних подій виконується ключова властивість, яка дозволяє отримати набагато більше практично значимих результатів, ніж для залежних подій: якщо дві події незалежні, то ймовірність їх сумісної (спільної) появи дорівнює добутку ймовірностей цих подій. Якщо одна подія – це правильна відповідь на і-те завдання тесту, а інша подія – правильна відповідь на j-те завдання тесту, то незалежність цих подій означає, що ймовірність правильної відповіді на обидва завдання одночасно дорівнює добутку окремих ймовірностей відповісти на кожне з цих завдань. Це ж саме стосується протилежних подій – неправильних відповідей на два завдання,а також їх комбінацій, коли одна подія означає правильну відповідь на одне з завдань, а інша подія – неправильну відповідь на інше завдання.

Локальна незалежність, виконання якої вимагається в IRT, означає, що відповіді на завдання тесту як події є незалежними для будь-якої гомогенної підпопуляції осіб. Термін «локальна» походить від того факту, що гомогенній субпопуляції осіб відповідає одна точка на осі латентної характеристики.

Вимога *одномірності* означає, що статистична залежність між завданнями може бути пояснена єдиною латентною характеристикою. Тест буде одномірним, якщо його завдання є статистично *залежними* по всій популяції екзаменованих, і існує єдина латентна характеристика така, що завдання є статистично *незалежними* у кожній гомогенній субпопуляції з даної популяції.

Зауважимо, що тест може бути і двомірним, і більшої вимірності. Двомірність, наприклад, означатиме, що існують дві латентні характеристики такі, що для субпопуляції, гомогенної одночасно по обох них, виконується локальна незалежність. Таким чином, можна сказати, що розмірність тесту дорівнює кількості латентних характеристик, необхідних для досягнення локальної незалежності. Тут будемо розглядати лише одновимірну IRT. Необхідно чітко усвідомлювати шкоду, яку може завдати багатовимірність, якщо вона трактується як одновимірність. Проілюструємо це на такому прикладі. Обираючи майбутню професію, молода людина може міркувати так: лікарем бути краще, ніж токарем, тому що професія лікаря більш престижна;

токарем бути краще, ніж учителем, тому що токар отримує більшу заробітну платню; учителем бути краще, ніж лікарем, тому що я боюся вигляду крові. Отримали суперечливий ланцюжок переваг: лікар → токар → учитель → лікар. Суперечливість є наслідком багатовимірності критеріїв (престижність, заробітна платня, страх перед виглядом крові).

Однією із найбільших переваг IRT є те, що вона дозволяє порівнювати опитуваних, яким пред'являються не одні й ті ж завдання тесту. Таке вимірювання називають *вимірюванням, вільним від тесту*. Проілюструємо цю властивість на прикладі. Нехай тест складається з чотирьох завдань, характеристичні криві яких мають східчастий вигляд (рис. 9.3).

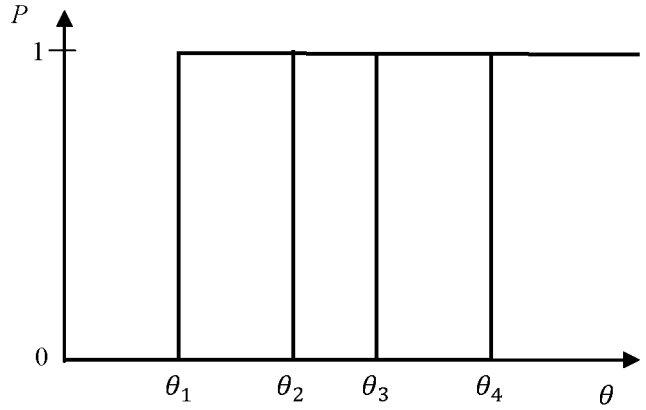

Рис. 9.3. Характеристичні криві 4-х завдань

Нехай завдання пронумеровані за рівнем трудності (зліва направо на рисунку). Припустимо, що перші два завдання пред'являлися особі А, а два останні завдання – особі Б. Нехай особа А відповіла на перше завдання правильно, а на друге – неправильно. Звідси робимо висновок, що рівень латентної характеристики у А знаходиться між θ_1 і θ_2. Нехай особа Б відповіла правильно на завдання 3 і неправильно на завдання 4. Тоді її рівень латентної характеристики знаходиться між θ_3 і θ_4. Отже, у особи Б рівень латентної характеристики вищий, ніж у особи А. Більш точно визначити положення осіб на прямій ми не можемо, оскіль-

ки їм пред'являлося замало завдань. Але коли завдань багато, положення особи на осі латентної характеристики можна визначити фактично як точку. Це зауваження є важливим тому, що ми могли б такі ж міркування навести для тестових завдань, трудність яких визначена апробацією у межах класичної теорії. У випадку, коли характеристичні криві завдань є S-подібними (що зазвичай і буває), міркування будуть аналогічними, хоча й не такими ж очевидними.

Логістичні моделі. Вище ми переконалися у тому, що залежність між рівнем латентної характеристики та ймовірністю правильної відповіді на завдання добре описується за допомогою моделі – S-подібної кривої. Також ми зазначали, що функція нормального розподілу є однією з таких кривих. Саме графік функції нормального розподілу (огіва) був домінуючою формою для кривих ICC у найбільш ранніх дослідженнях з IRT. Пізніше почали використовуватися так звані логістичні моделі, які, з одного боку, дозволяють спростити необхідні обчислення, а з іншого боку, можуть враховувати додаткові параметри тестових завдань, такі, як роздільна здатність та вплив ефекту вгадування.

Основою для всіх логістичних моделей є *кумулятивна логістична функція*. Її рівняння можна записати для i-го завдання як

$$P_i(\theta) = \frac{e^x}{1 + e^x},$$

де x – це змінна, пов'язана певним чином з θ.

В IRT розглядають три логістичні моделі, які відрізняються кількістю додаткових параметрів.

Однопараметрична логістична модель (часто позначається як 1PL) задається формулою

$$P_i(\theta) = \frac{e^{d(\theta - b_i)}}{1 + e^{1,7(\theta - b_i)}}.$$

Тут параметр b_i відповідає за *трудність завдання* – поняття аналогічне до такого у класичній теорії. Для різних завдань значення

цього параметра є різним. При $d = 1{,}7$ крива є максимально близькою до функції нормального розподілу.

При $d = 1$ однопараметрична модель IRT еквівалентна моделі датського математика Георга Раша, яку той використовував у своїй теорії вимірювання латентних змінних, що базується на відмінних від IRT концепціях і має назву Rasch Measurement – теорія вимірювань Раша.

Усі криві, які описуються однопараметричною моделлю, відрізняються одна від одної при різних значеннях параметра b_i лише зсувом вздовж осі θ, їх кривизна залишається незмінною.

У *двопараметричній моделі Бірнбаума* (2PL) вводиться додатковий параметр a_i:

$$P_i(\theta) = \frac{e^{1,7a_i(\theta - b_i)}}{1 + e^{1,7a_i(\theta - b_i)}}.$$

Параметр a_i входить, на відміну від b_i, як множник до аргументу θ, тому зміна значення цього параметра призводить до зміни кривизни кривої. Тому зміст цього параметра можна інтерпретувати як роздільну здатність завдання. На рисунку 9.4 зображено дві криві, які відрізняються значенням параметра a_i: $a_i' > a_i''$.

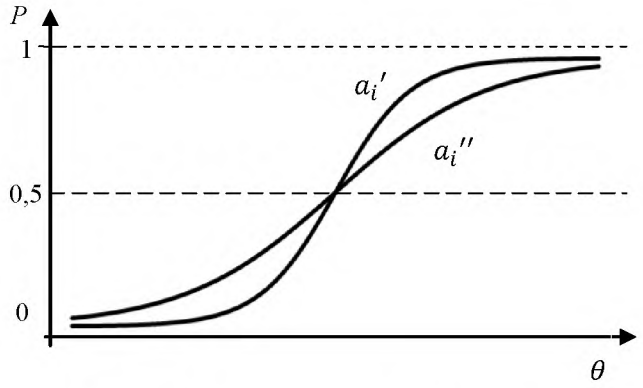

Рис. 9.4. Вплив параметра a_i на форму ICC

Крива з більшим значенням параметра є більш крутою в середній області. Це означає більшу роздільну здатність завдання для осіб, чий рівень близький до середнього відносно даного завдання. Справді, зміна рівня латентної характеристики у цій області на крок $\Delta\theta$ веде до більшої зміни ймовірності для завдання з більшим значенням a_i. Але слід пам'ятати, що на кінцях області θ картина протилежна: крива з більшим значенням параметра у цих областях більш полога, а отже, й роздільна здатність завдання менша.

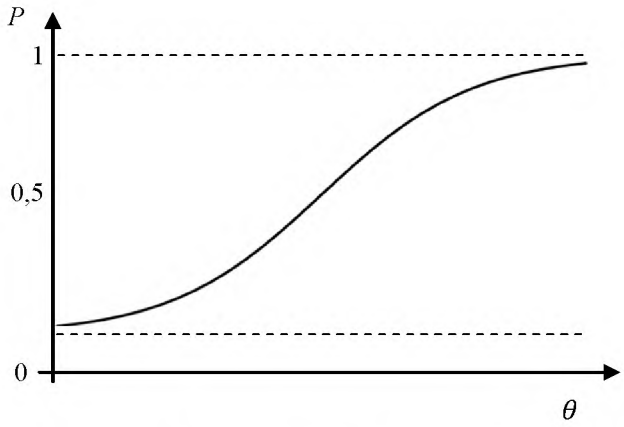

Рис. 9.5. Крива 3PL

Трипараметрична логістична модель Бірнбаума (3PL) містить ще один параметр, який відображає вплив ефекту вгадування на ймовірність відповіді на завдання. Як видно з малюнку 9,4, криві одно- і двопараметричної моделей наближаються до горизонтальної осі при зменшенні θ до мінімального значення, тобто ймовірність правильної відповіді для осіб з мінімальним рівнем латентної характеристики наближається до нуля.

Але якщо завдання має форму множинного вибору (а саме такими зазвичай і є завдання з дихотомічною оцінкою, які тут розглядаються), і передбачається, що екзаменований, котрий не знає правильної відповіді, намагатиметься її вгадати, то у нього вже є гарантована ймовірність, яка залежить від кількості варіантів відповіді. Так, якщо варіантів відповіді є 4, то у екзаменованого з

найнижчим рівнем вже є ймовірність 0,25 відповісти на завдання правильно. Тобто характеристична крива такого завдання повинна наближатися на лівому кінці до значення 0,25, а не до нуля (рисунок 9.5). За це відповідає у трипараметричній моделі величина параметра c_i. Рівняння для кривої ICC трипараметричної моделі має вигляд:

$$P_i(\theta) = c_i + \frac{(1 - c_i)e^{1,7a_i(\theta - b_i)}}{1 + e^{1,7a_i(\theta - b_i)}}.$$

Для завдання множинного вибору з чотирма варіантами відповіді $c_i = 0,25$.

Шкала вимірювань. Шкала латентної характеристики може мати будь-який початок і одиницю вимірювання. Зазвичай їх підбирають так, щоб середнє значення латентної характеристики дорівнювало нулю, а її стандартне відхилення – одиниці для цільової популяції. Отримана таким чином шкала буде мати як додатні, так і від'ємні значення. Значення параметрів обраної для кожного завдання моделі залежать від обраної шкали. Шкала допускає будь-які лінійні перетворення виду

$$\theta' = k\theta + l,$$

де k і l можуть бути довільними числами. Тоді параметри моделі b_i та a_i потрібно перетворити так:

$$b'_i = kb_i + l,$$

$$a'_i = \frac{a_i}{k}.$$

Процедури оцінювання параметрів моделі. Параметри моделі, обраної для даного завдання, мають бути оцінені. Для цього існують щонайменше дві загальноприйняті ітеративні процедури. Одна з них заснована на відомому статистичному методі *максимальної правдоподібності*, іншу називають *евристичною, чи апроксимаційною процедурою*. Суть цих процедур, відповідні фо-

рмули і алгоритми ми не описуємо в цій книзі, вони повинні бути предметом спеціального курсу з IRT. Зазвичай саму процедуру виконує комп'ютерна програма, оскільки вона містить дуже велику кількість обчислень. Тут перелічимо лише варіанти реалізації методу максимальної правдоподібності.

Сумісна процедура максимальної правдоподібності дозволяє шукати одночасно як параметри завдань, так і рівні латентної характеристики екзаменованих. Ця процедура підходить для всіх трьох логістичних моделей, але для оцінки параметрів трипараметричних моделей вона вимагає значних за об'ємом вибірок екзаменованих. Немає прямих способів перевірити слушність (збіжність за ймовірністю до істинних значень) знайдених оцінок параметрів, єдиний доказ слушності отримаємо, якщо вдасться довести, що оцінки параметрів збігаються до своїх істинних значень із збільшенням об'єму вибірки.

Інша процедура називається *методом маргінальної максимальної правдоподібності*. Основною перевагою цього методу є можливість доведення слушності отриманих оцінок параметрів.

Ще одна процедура називається *умовною процедурою оцінки максимальної правдоподібності*. Оцінки, отримані за цією процедурою, є слушними, і це є основною її перевагою.

Розглянемо далі два важливих застосування теорії IRT: калібрування тестових завдань та комп'ютерне адаптивне тестування.

Калібрування завдань. Значним внеском в теорію і практику аналізу тестових завдань було би виявлення таких параметрів, які були б відносно інваріантними щодо змін у якісному складі екзаменованих. Якщо параметри завдань є інваріантними, то можна їх оцінити на основі однієї групи осіб, а потім впевнено застосовувати для будь-якої іншої групи осіб.

Класичний аналіз характеристик завдання, таких як частка правильних відповідей на завдання чи кореляція між завданням та тестом, не є інваріантними щодо вибору екзаменованих. Натомість, багато досліджень свідчать на користь інваріантності параметрів логістичних моделей IRT.

Інваріантність дає змогу оцінювати параметри множини завдань навіть за умови, коли кожен екзаменований відповідає лише на частину завдань цієї множини. Ця властивість називається *калібруванням завдання, не залежним від екзаменованого*.

Припустимо, що потрібно оцінити параметри 75 завдань, пред'являючи кожному учаснику апробації по 50 завдань. Для цього ми можемо розбити всі завдання на частини А, Б і В, по 25 завдань у кожній, і пред'явити одній групі частини А і Б, другій – А і В. Спільна для всіх учасників частина А використовується для створення спільної шкали, на якій можуть бути поміщені усі оцінки параметрів інших завдань. Нехай для описання всіх 75 завдань було обрано двопараметричну логістичну модель. Для кожної з груп учасників шкала латентної характеристики утворюється так, щоб середнє значення дорівнювало нулю, а стандартне відхилення дорівнювало одиниці. Нагадаємо, що значення параметра трудності завдань b_i визначається як точка на шкалі латентної характеристики, для якої $P_i(\theta) = 0{,}5$. Значення параметра b_i, обчислені для кожної групи, поміщаються на шкалу цієї групи. Таким чином, значення параметра для завдань частини Б опиняються на шкалі першої групи, а для завдань частини В – на шкалі другої групи. Для частини ж завдань А маємо два набори оцінок параметра, по одному набору для кожної групи учасників. Один набір виражений у шкалі оцінок першої групи, інший – у шкалі другої групи. Нехай ми вирішили помістити на шкалу для першої групи параметри завдань частини С, яких там не вистачає для повного комплекту завдань тесту. Оскільки ми можемо застосовувати лінійне перетворення шкали

$$b_i' = kb_i + l$$

для того, щоб оцінки параметра b_i, отримані на шкалі другої групи, перевести в оцінки b_i' на шкалі першої групи, то проблема полягає у відшуканні прийнятних значень k і l. Так само значення k потрібне для перетворення параметра роздільної здатності a_i завдань частини С зі шкали другої групи у шкалу першої групи за допомогою перетворення

$$a_i' = \frac{a_i}{k}.$$

Тут і стають у пригоді оцінки параметрів трудності завдань, знайдені для спільної частини А в обох шкалах. Адже вони повин-

ні бути зв'язані тим самим співвідношенням, що й оцінки параметрів частини С. Якщо $b_i^{A_1}$ – оцінки параметра трудності завдань частини А, виражені у першій шкалі, а $b_i^{A_2}$ – ці ж оцінки, виражені у другій шкалі, то

$$b_i^{A_1} = kb_i^{A_2} + l,$$

тобто оцінка на першій шкалі є лінійною функцією оцінки на другій шкалі, точки з відповідними координатами лежать на прямій лінії, для якої k є тангенсом кута нахилу, а l - точкою перетину з вертикальною віссю.

Практичне значення калібрування завдань очевидне. Маючи банк таких завдань, який може наповнюватися поступово, шляхом пред'явлення новим групам з цільової популяції нових форм тесту, які містять частину вже апробованих завдань (таку частину називають *якірною*), ми можемо надалі пред'являти різним групам у різний час різні форми тесту, маючи при цьому змогу поміщати екзаменованих на єдину шкалу латентної характеристики. Цим самим ми позбудемося проблеми небажаного повторення завдань у тестах, які пред'являються у різний час, і тому стають відомими для тих, хто екзаменується у другу чергу.

Комп'ютерне адаптивне тестування. Ще одним цікавим застосуванням теорії IRT є комп'ютерне адаптивне тестування (Computerized Adaptive Testing, CAT). Ідея адаптивного тестування з'явилася унаслідок універсального недоліку, властивого тестам, які призначені для вимірювання латентної характеристики у популяції осіб з широким діапазоном її мінливості. Наприклад, стандартизований тест навчальних досягнень, більшість завдань якого, як і рекомендує теорія, мають середню для даної популяції трудність, погано диференціює найбільш слабких і найбільш сильних представників цільової популяції. Очевидно, що найбільш слабкі особи не зможуть відповісти правильно на більшість завдань, і тому результати тестування всередині слабкої субпопуляції будуть надто схожими між собою і тому погано диференціюватимуть осіб з цієї субпопуляції. Так само, на більшість завдань тесту представники найбільш сильної субпопуляції відповідатимуть правильно, і їх результати будуть надто схожими і часто збігатимуться, що не

дозволить розрізнити більш сильного представника від більш слабкого у цій сильній субпопуляції. Оскільки тест складається з окремих завдань, то ця ж проблема спостерігається і для окремого тестового завдання. Очевидно, що на дане завдання середньої трудності два найбільш слабкі учасники відповідатимуть однаково неправильно з великою ймовірністю, а два найбільш сильні учасники відповідатимуть правильно знову ж таки з великою ймовірністю. Цю проблему добре ілюструє S-подібна форма графіка функції відповідей на завдання (кривої ICC для дихотомічного завдання): її середня частина є набагато більш крутою, ніж кінці, а це означає, що із зміною рівня латентної характеристики на одиницю ймовірність правильної відповіді найбільше змінюється для осіб з близьким до середнього рівнем, і найменше змінюється для осіб з дуже високим або дуже низьким рівнем. Власне, це й характеризується роздільною здатністю завдання, але, як показує двопараметрична логістична модель, роздільна здатність завдання є різною для різних θ, і чим більшою вона є для близьких до середніх значень θ, тим меншою вона є для екстремальних значень, тобто не існує завдань, які б однаково добре диференціювали осіб на всьому діапазоні мінливості латентної характеристики в популяції.

Сказане можна резюмувати простим інтуїтивно зрозумілим твердженням: тестувати слабких потрібно легким тестом, а тестувати сильних потрібно складним тестом. Або, більш точно: для кожної субпопуляції з однаковим рівнем латентної характеристики повинен існувати свій тест, з завданнями, які мають середню трудність для цієї субпопуляції.

Але тоді виникає інша проблема: на початку тестування рівень особи невідомий, власне, тестування й покликане визначити цей рівень. Як пред'явити цій особі ідеальний для неї за рівнем трудності тест? Принцип вирішення цієї проблеми у комп'ютерному адаптивному тестування наступний. Комп'ютер пред'являє екзаменованому завдання середньої для популяції трудності. Якщо той відповідає правильно, комп'ютер вибирає з банку відкаліброваних завдань більш складне завдання і пред'являє його екзаменованому. На кожному наступному кроці комп'ютер визначає рівень екзаменованого, виходячи з його відповідей на всі пред'явлені на попередніх кроках завдання, і підбирає з банку завдань чергове завдання, яке найкраще підходить для цього рівня.

Таким чином, комп'ютерне адаптивне тестування полягає у поступовому, ітеративному уточненні рівня екзаменованого найбільш швидким та ефективним шляхом. Умовою зупинки процедури є досягнення заданої точності вимірювання. Процедура під силу лише комп'ютеру, оскільки визначення на кожному кроці досягнутого екзаменованим рівня та вибір найбільш придатного для цього рівня чергового завдання вимагає великої кількості досить складних обчислень, які повинні відбуватися швидко, в реальному часі.

Тепер подивимося, як адаптивне тестування реалізується у рамках теорії IRT.

Описуючи проблему, ми оперували поняттям роздільної здатності завдання. З цим поняттям тісно пов'язані інші важливі поняття IRT: *інформаційної функції завдання* та *інформаційної функції тесту*.

Роздільна здатність завдання у кожній точці латентної характеристики виражається крутизною графіка функції відповідей на завдання, тобто кутом нахилу дотичної до кривої цієї функції у даній точці. Кут нахилу дотичної до графіка деякої функції у точці, точніше, тангенс цього кута – це число, яке є похідною даної функції у цій точці. Сукупність похідних у всіх точках області визначення функції – це функція, яка є похідною даної функції.

Похідна функції відповідей на завдання визначає форму *інформаційної функції* (або, простіше, *інформації*) *i*-го завдання:

$$I_i(\theta) = \frac{\left(P_i'(\theta)\right)^2}{P_i(\theta)Q_i(\theta)}.$$

Тут у чисельнику знаходиться квадрат похідної функції відповідей на завдання, а $Q_i(\theta) = 1 - P_i(\theta)$.

На рисунку 9.6 зображено графік інформаційної функції завдання поруч з кривою функції відповіді на завдання.

Як бачимо, інформація є максимальною у точці перегину кривої функції відповіді на завдання, тобто у тій точці θ', для якої трудність (ймовірність правильної відповіді для дихотомічного завдання) дорівнює 0,5.

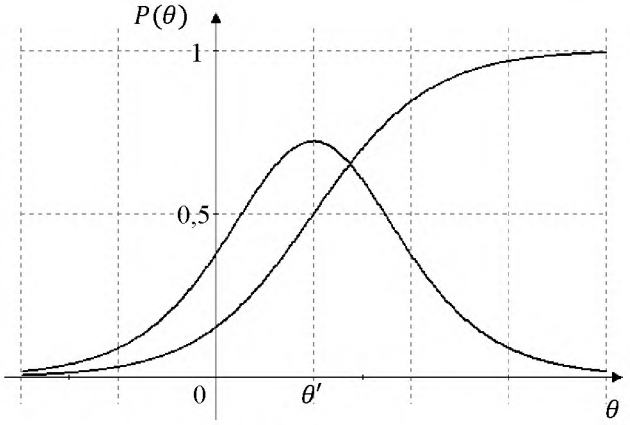

Рис. 9.6. ICC та інформаційна функція

Вище ми зазначили, що на кожному кроці процедури адаптивного тестування комп'ютер визначає рівень опитуваного на підставі інформації про його відповіді на попередні завдання. Отже, комп'ютер повинен якимось чином узагальнити інформацію про отримані від опитуваного відповіді у вигляді поточної оцінки рівня латентної характеристики .

Один із способів оцінювання θ опитуваного базується на методі максимальної правдоподібності. Функція правдоподібності – умовна ймовірність того, що за даного набору завдань з відомими параметрами їх логістичних моделей (позначимо всю множину параметрів через β), та відомому рівню θ опитуваного, буде отриманий вектор відповідей x (для дихотомічних завдань це набір нулів за правильні відповіді та одиниць – за неправильні):

$$P(x|\theta,\beta) = \prod_i P_i(\theta)^{x_i} Q(\theta)^{1-x_i},$$

де $Q(\theta) = 1 - P(\theta)$. У правій частині функції стоїть добуток по всіх завданнях таких функцій: якщо відповідь на якесь завдання правильна, то це функція відповіді на завдання $P(\theta)$, якщо відповідь неправильна, то це функція $Q(\theta)$. Тоді рівень опитуваного –

це точка на осі латентної характеристики, для якої функція у правій частині формули має максимум.

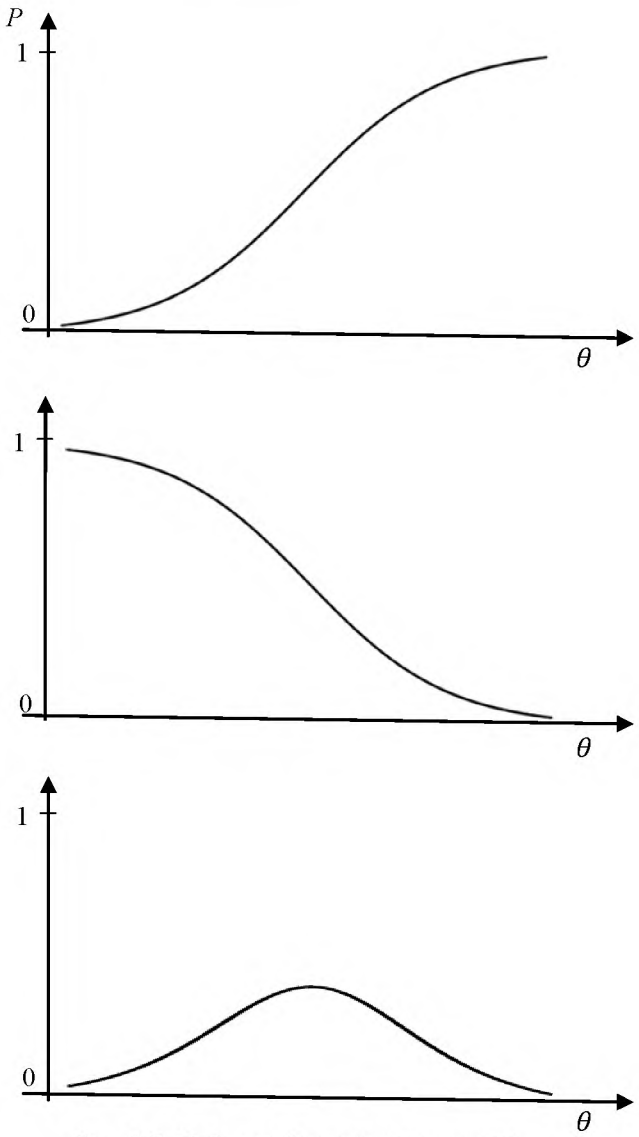

Рис. 9.7. Добуток функцій правильної і неправильної відповідей

Нехай, наприклад, опитуваний відповідав на два завдання, і дав на перше завдання правильну відповідь, а на друге – неправильну. На рисунку 9.7 зображено: вгорі – функцію ймовірності правильної відповіді на перше завдання, посередині – функцію ймовірності неправильної відповіді на друге завдання, внизу – добуток цих двох функцій. Рівень опитуваного θ – це та точка, у якій крива унизу рисунка має максимум.

Знайти максимум функції правдоподібності можна чисельними методами. Для цього замість самої функції розглядають її логарифм – *логарифмічну функцію правдоподібності*. Це спрощує подальші обчислення, оскільки логарифм добутку дорівнює сумі логарифмів:

$$\ln P(x|\theta,\beta) = \sum_i (x_i \ln P(\theta) + (1 - x_i) \ln Q(\theta))$$

Логарифмічна функція правдоподібності має максимум у тій же точці, що й сама функція правдоподібності. Для відшукання максимуму потрібно взяти похідну логарифмічної функції правдоподібності по змінній θ

$$\ln' P(x|\theta,\beta) = \sum_i (x_i - P_i(\theta)) \frac{P'_i(\theta)}{P_i(\theta)Q_i(\theta)},$$

прирівняти її до нуля, і розв'язати отримане рівняння відносно θ, що можна зробити методом Ньютона.

Більш розвинений спосіб оцінки рівня опитуваного за відповідями на отримані завдання – *Байєсова модальна оцінка*. Ця оцінка базується на апостеріорному розподілі

$$p(\theta|x) \propto L(\theta|x)p(\theta),$$

де $p(\theta)$ – деяка апріорна інформація про θ. Цю апріорну інформацію можна сприймати як наслідок додавання до набору отриманих опитуваним завдань додаткового завдання. Якщо апріорна інформація однакова для кожного значення θ, то це не додає нічого, і апостеріорний розподіл θ буде пропорційним до функції правдо-

подібності. На рис. 9.8 зображена ситуація, коли апріорною інформацією є нормальний розподіл θ, і опитуваний отримав два завдання, на які відповів правильно. Внизу рисунка знаходиться функція-добуток цих трьох функцій, її максимум вказує на значення латентної характеристики θ опитуваного на даний момент.

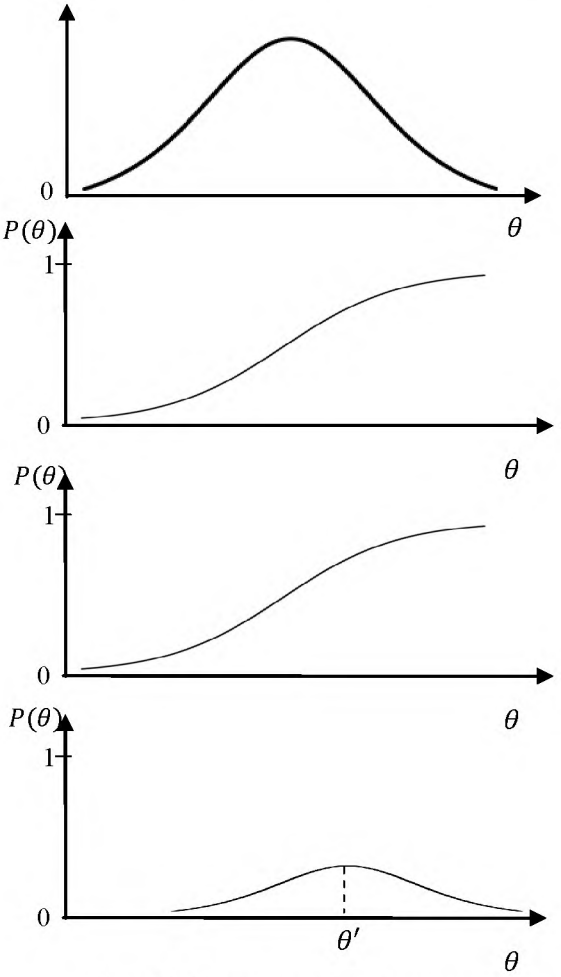

Рис. 9.8. Схематичне представлення Байєсової оцінки

Відшукання точки рівня опитуваного, яка відповідає максимуму Байєсової модальної оцінки, аналогічне до випадку використання функції правдоподібності.

Основні проблеми, з якими може стикатися дослідник при оцінюванні рівня латентної характеристики опитуваного на основі його відповідей на отримані завдання описаними вище методами – це порушення унімодальності розподілу оцінки і випадки, коли опитуваний відповідає на всі завдання однаково правильно або неправильно. При порушенні унімодальності крива інтегральної оцінки (функція правдоподібності чи Байєсова модальна оцінка) має кілька локальних максимумів, тобто похідна дорівнює нулю у кількох точках, і це утруднює пошук глобального максимуму функції. У цьому випадку велику роль відіграє вдалий вибір початкової точки ітеративного процесу пошуку розв'язку. У випадку ж, коли всі відповіді опитуваного є однаково правильними або неправильними, оцінка його рівня за функцією правдоподібності дорівнює, відповідно, плюс або мінус нескінченності, і тоді слід долучати до процесу оцінювання додаткову апріорну інформацію, тобто використовувати Байєсову модальну оцінку.

Раніше ми зазначили, що процес уточнення рівня опитуваного є ітеративним наближенням до істинного значення шляхом пред'явлення йому нових завдань. Якщо цей процес є дійсно збіжним, то правилом зупинки алгоритму може бути досягнення різницею між попереднім і наступним значеннями оцінки достатньо малої заданої наперед величини. Але наскільки можна довіряти знайденому остаточному значенню? Оцінити дисперсію знайденої оцінки у випадку, коли набір отриманих опитуваним завдань є достатньо великим, можна наближено як обернене значення *інформаційної функції тесту*. Це функція, яка є простою сумою інформаційних функцій завдань:

$$I(\theta) = \sum_i \frac{\left(P_i'(\theta)\right)^2}{P_i(\theta)Q_i(\theta)}.$$

Вона на залежить від відповідей конкретного опитаного, а лише від самого набору отриманих ним завдань, і є адитивною.

Потрібно розрізняти інформацію, забезпеченою формулою оцінювання, від інформації, забезпеченої тестом. Інформація по тесту є верхньою межею для інформації, отриманої від будь-якого частинного випадку оцінки.

Найпростіша схема оцінювання полягає у присвоюванні одного балу за кожну правильну відповідь і нуля балів – за кожну неправильну відповідь. Загальна оцінка має вигляд

$$X = \sum_i U_i,$$

де U_i набуває значення, відповідно, 0 або 1. Для однопараметричної моделі ця схема забезпечує максимально можливе значення інформаційної функції оцінки.

Для двопараметричної моделі максимальне значення інформаційної функції оцінки дає зважена сума $X = \sum_i a_i U_i$.

Порівняння теорії IRT та класичної теорії тестування. Відмінності теорії IRT від класичної тестової теорії полягають у:

- природі та деталях початкових припущень;
- більшій зосередженості на окремих завданнях тесту як незалежних його елементах;
- більшій увазі до результатів окремих опитуваних, ніж до загальних результатів груп опитуваних;
- використанні різноманітних шкал чи метрик поза рамками первинних балів;
- типами, широтою, та глибиною передбачення;
- важливістю перевірки точності використовуваних моделей та передбачень.

Класична тестова теорія використовує відносно прості означення та широкі припущення: спостережена оцінка опитуваного визначається як сума його істинної оцінки і похибки вимірювання, похибки оголошуються некорельованими з істинними оцінками та іншими похибками. Ці припущення дозволяють отримувати результати, пов'язані перш за все з властивостями спостережених оцінок, такі як надійність чи стандартна похибка вимірювання. Класична теорія розглядає статистичні властивості тестових завдань, такі як трудність завдання, які повністю залежать від груп

опитуваних, для яких ці властивості були отримані, і не дає засобів для узагальнення цих властивостей на інші групи опитуваних. Класична теорія оперує передусім з первинними балами за тест як сумою балів, отриманих за правильні відповіді, і не пропонує засобів для врахування ефекту зміни метрики або зміни завдання у тесті.

Припущення IRT є більш жорсткими, ці пропущення стосуються того, як окремий опитуваний з певним рівнем латентної характеристики відповідатиме на окреме завдання. Моделі зв'язків між оцінками за завдання, рівнями опитуваних, та характеристиками завдань є в IRT нелінійними. IRT дозволяє робити як безумовні (для груп опитуваних) так і умовні (для опитуваних з визначеним рівнем латентної характеристики) детальні передбачення. Механізм передбачення, у порівнянні з класичною теорією, є більш гнучким і дає готові результати у одиницях вимірювання відмінних від первинної суми балів.

Чому ж використання IRT досі не набуло масового поширення? Тому що для моделювання завдань тесту методами цієї теорії потрібні значно більші вибірки з цільової популяції; вона вимагає складних обчислень з використанням комп'ютерів. Не менш важливою з практичної точки зору є неочевидність результатів тестування, отриманих методами IRT, трудність пояснення цих результатів як самим опитуваним, так і іншим стейкхолдерам (групам зацікавлених осіб). Останнє є особливо важливим у випадку тестування високої відповідальності, яким є, скажімо, зовнішнє незалежне тестування випускників школи в Україні.