8. АНАЛІЗ ТЕСТОВИХ ЗАВДАНЬ

Одним із ключових етапів конструювання стандартизованого тесту є їх аналіз його завдань, заснований на даних, отриманих під час польового дослідження тесту (тобто апробації тесту на великій репрезентативній вибірці з цільової популяції). Числові характеристики тестових завдань, отримані на основі даних польового дослідження, називають ще *психометричними характеристиками завдань*. Ці характеристики часто зберігають разом із завданнями у банку завдань, щоб потім підібрати такі завдання для конкретного тесту, які забезпечать потрібну його якість. Деякі характеристики завдань отримують в рамках класичної теорії тестування, описаної нами в 5 главі. Але також для широкомасштабних вимірювань з великою цільовою популяцією використовують і так звану сучасну теорію тестування – так у нас часом перекладають англійську назву теорії *Item Response Theory* (IRT), яка не може бути перекладеною українською дослівно. Ця теорія значною мірою орієнтована саме на аналіз окремих тестових завдань, і далі ми познайомимося з основами цієї теорії.

Трудність завдань. Для завдання, яке оцінюється за дихотомічною шкалою «правильно-неправильно» (будемо надалі називати їх просто дихотомічними завданнями), *трудність* (або *складність*) завдання визначається як відносна частка тих екзаменованих, які відповіли на завдання правильно. Нехай у репрезентативній вибірці 100 осіб, і 60 з них правильно відповіли на дане завдання. Тоді трудність завдання

$$p = \frac{60}{100} = 0{,}6.$$

Ми не даремно позначили цю характеристику буквою p - тією ж, що ймовірність, адже отримана величина – це приблизна ймовірність того, що представник цільової популяції відповість на завдання правильно. Так склалося, що терміном «трудність» ми насправді позначаємо легкість завдання: чим більша трудність

завдання у визначеному нами технічному сенсі, тобто чим більше p, тим легше на нього відповісти екзаменованому. Очевидно, що ймовірність відповісти на завдання *неправильно* дорівнює, як ймовірність протилежної події,

$$q = 1 - p.$$

Якщо тест цілком складається з дихотомічних завдань, то його трудністю є сумарна трудність завдань $\sum_i p_i$. Також вживають таку характеристику як *середня трудність завдань тесту* – це середнє арифметичне трудностей усіх завдань, які входять до тесту. Для тесту з k завданнями середня трудність завдань дорівнює

$$\bar{p} = \frac{\sum_{i=1}^{k} p_i}{k}.$$

Нехай на одне дихотомічне завдання з трудністю p відповідала група з n осіб. Ця ситуація відповідає відомій у теорії ймовірностей схемі Бернуллі – повторне проведення n незалежних випробувань з ймовірністю «успіху», яка у кожному випробуванні дорівнює p. Кількість правильних відповідей (тобто кількість успіхів) у цій групі має біноміальний розподіл з параметрами n і p. Зокрема, ймовірність того, що правильну відповідь дадуть рівно k з n осіб, обчислюється за формулою Бернуллі:

$$P_n(k) = C_n^k p^k q^{n-k}.$$

Ця величина із зростанням k спочатку зростає, потім, досягши максимуму спадає. Якщо $(n+1)p$ – ціле число, то максимально можливими є дві кількості вгадувань – саме число $(n+1)p$, та на одиницю менше. Так, якщо на завдання з трудністю 0,25 відповідали 19 осіб, то, найбільш імовірно, правильну відповідь дали $(19+1) \times 0{,}25 = 5$ осіб або 4 особи. Якщо $(n+1)p$ не є цілим числом, то максимально можливою кількістю правильних відповідей буде ціла частина цього числа. Наприклад, якщо на завдання з трудністю 0,25 відповідали 20 представників цільової популяції, то найбільш імовірно, правильних відповідей буде $[(20+1) \times 0{,}25] = [5{,}25] = 5$. Сама ймовірність цієї найбільш можливої кіль-

кості обчислюється за формулою Бернуллі, або, у випадку великої кількості екзаменованих, за асимптотичними (приблизними) формулами.

Дисперсія оцінок за дихотомічне завдання із трудністю p у групі осіб, тобто випадкової величини, яка представлена двома значеннями 0 (за правильну відповідь) та 1 (за неправильну відповідь) дорівнює pq. Це важливий факт для розуміння, чому завдання з трудністю 0,5 вважається найкращим для тестування. Для такого завдання дисперсія дорівнює 0,25. В усіх інших випадках дисперсія буде меншою. Тобто мінливість відповідей у групі осіб, які відповідали на одне дихотомічне завдання, буде максимальною у випадку середньої його трудності. Проілюструємо це на прикладах. Нехай у групі екзаменованих 10 осіб. Якщо завдання має середню трудність, то найбільш імовірно, що 5 осіб дадуть правильну відповідь і 5 осіб – неправильну. Це дозволяє виділити $5 \times 5 = 25$ пар осіб, у кожній з яких можна виділити бкльш сильного та більш слабкого екзаменованого. Для порівняння, нехай ця група відповідала на завдання з трудністю 0,1. Тоді в середньому лише одна з цих осіб відповість правильно, а інші 9 неправильно, і ми можемо скласти лише 9 пар осіб, у яких екзаменовані розрізнятимуться за успішністю.

Сказане не означає, що тест повинен складатися лише з завдань середньої трудності, адже такий тест погано диференціюватиме найбільш сильних та найбільш слабких екзаменованих. Завдання середньої та близької до неї трудності повинні переважати, але має бути невелика кількість складніших та легших завдань.

Поняття трудності *політомічного* завдання, тобто такого завдання, відповідь на яке може бути не тільки повністю правильною чи неправильною, але й частково правильною, в рамках класичної теорії тестування визначити важко. Ці завдання можуть належати до одного з двох типів. Перший тип – це завдання, яке складається з послідовності кроків, кожен з яких оцінюється окремо за дихотомічною шкалою. Якщо екзаменований на якомусь кроці дає неправильну відповідь, це тягне за собою неправильні відповіді й на усіх наступних кроках. Прикладом такого завдання може бути завдання спростити математичний вираз, якщо ця процедура вимагає кількох послідовних алгебраїчних дій. Другий тип – це завдання з кількома не пов'язаними одна з одною правильни-

ми відповідями, наприклад, серед множини міст вибрати ті, які є столицями держав.

Трудність дихотомічних завдань та ефект вгадування. У більшості випадків дихотомічні завдання мають форму завдання множинного вибору з однією правильною відповіддю. Таке завдання має той істотний недолік, що екзаменований, не знаючи правильної відповіді, може спробувати вгадати її, обираючи варіант відповіді навмання. Якщо, наприклад, завдання має чотири варіанти відповіді, то ймовірність чистого вгадування дорівнює ¼, тобто 0,25. Враховуючи, що вибір відповіді може здійснюватися екзаменованим не зовсім навмання, а з врахуванням інформації, що міститься у варіантах і може наштовхнути на відкидання деяких дистракторів, реальна ймовірність вгадування може бути ще більшою. Найгірша ситуація виникає, коли завдання має форму альтернативного вибору, тобто має лише два варіанти відповіді. У цьому випадку ймовірність чистого вгадування дорівнює 0,5. Якщо дати непосильне для даної групи екзаменованих завдання альтернативної форми, і дозволити їм вгадувати, то близько половини осіб відповість на нього правильно, і в результаті отримаємо «трудність» $p = 0{,}5$. Очевидно, не можна вважати непосильне завдання завданням середньої трудності, як це випливає з даного нами означення трудності дихотомічних завдань. Тому слід розуміти, що в означення ми закладали відсутність ефекту вгадування. Що відбудеться, коли ми репрезентативній групі осіб пред'явимо завдання, трудність якого нам заздалегідь відома з іншого дослідження, здійсненого для даної цільової популяції? Нехай, наприклад, відомо, що істинна трудність завдання альтернативної форми становить 0,5. Якщо є підстави вважати, що екзаменовані у даній групі схильні до зловживання вгадуванням, то отримаємо таку ситуацію: близько половини екзаменованих дадуть правильну відповідь, бо знають її; з тієї половини, які не знають відповіді, половина вгадає її. Звідси отримаємо відсоток правильних відповідей: $50\% + 50\%/2 = 75\%$. Таким чином, «спостережена» трудність завдання дорівнюватиме 0,75. Для найбільш розповсюджених варіантів кількості відповідей у завданнях множинного вибору з однією правильною відповіддю для завдання середньої трудності значення спостереженої трудності подано у таблиці 8.1.

Таблиця 8.1. Спостережена трудність завдань множинного вибору з істинною трудністю 0,5 для різних випадків кількості варіантів відповідей

Кількість варіантів відповіді	2	3	4	5
Спостережена трудність	0,75	0,67	62,5	0,60

Емпіричні дослідження (Лорд) показують, що спостережена трудність на практиці є істотно вищою через те, що екзаменовані, які не знають правильної відповіді, все ж намагаються осмислювати інформацію, закладену у варіантах. Так, для завдання з істинною трудністю 0,5 і чотирма варіантами відповіді спостережена трудність на практиці складає близько 0,74. Розглянемо для прикладу завдання:

$$28 \times 7 = \cdots$$

а) 186
б) 196
в) 287
г) 554

Якщо екзаменований не може перемножити числа, він усе ж може знати, що $8 \times 7 = 56$, тому відповідь повинна закінчуватися цифрою 6. Для такого екзаменованого ймовірність вгадування вже буде не 0,25, а 0,5.

З практичної точки зору більш цікавою для нас є обернена задача: як у ситуації, коли допускається вгадування, знайти істинну трудність завдання за спостереженою трудністю? Для цього можна скористатися формулою, виведення якої рекомендується читачеві як вправа:

$$p_{\text{іст.}} = \frac{kp_{\text{спост.}} - 1}{k - 1},$$

де k – кількість варіантів відповіді у завданні. Наприклад, якщо $k = 5$, і правильно відповіли на завдання 60% екзаменованих, то спостережена трудність дорівнює 0,6, а істинна трудність завдання дорівнює

$$p_{\text{іст.}} = \frac{5 \times 0{,}6 - 1}{5 - 1} = 0{,}5.$$

Аналіз дистракторів методом порогових груп. Із наведеного вище поняття трудності завдання випливає, що використання у тестах завдань множинного вибору з однією правильною відповіддю є дуже зручним для аналізу. З іншого боку, приклад про добуток двох чисел ілюструє необхідність якісного конструювання дистракторів. Проблема полягає у тому, що аналіз однієї лише трудності завдання не дозволяє розгледіти погану якість дистракторів, оскільки остання формула годиться лише для врахування чистого вгадування, тобто вгадування навмання. Якщо дистрактори були неякісними, отримаємо неправильну оцінку істинної трудності завдання. Тому дуже важливо уміти на основі результатів апробації тесту проаналізувати якість дистракторів окремого завдання. Один із математичних способів розгледіти проблему у дистракторах – так званий метод аналізу порогових груп. Відразу зауважимо, що цей метод вимагає тестування у повному об'ємі, а не лише у вигляді пред'явлення екзаменованим одного досліджуваного завдання. Інформація про виконання усього тесту, за умови задовільної валідності його завдань, дозволяє диференціювати екзаменованих за рівнем вимірюваної якості. Тоді можна поділити усю групу екзаменованих на підгрупи за цим рівнем. Покладемо для визначеності, що вирішено розглядати 5 підгруп, рівних за кількістю учасників. Тоді потрібно лише ранжувати учасників у порядку зростання їх успішності і поділити отриманий список на 5 приблизно рівних частин. Таким чином, у першу підгрупу потрапляє 20% тих, хто справився з тестом найгірше, у другу – 20% дещо сильніших, і так далі. Для кожної з утворених підгруп визначається, який відсоток її членів обирав той чи інший варіант відповіді. Далі для полегшення аналізу дані візуалізують у вигляді діаграми.

Розглянемо для прикладу аналіз завдання, яке пропонувалося у пробному інтернет-тестуванні з математики, розміщеному на сайті pitest.org.ua (результати аналізу люб'язно надані автору розробником тесту А. Миляником).

Завдання 13. Розв'яжіть нерівність $|x-3| \leq 1$.

А	Б	В	Г	Д
$(-\infty, 4]$	$[-4, 2]$	$[-4, -2]$	$[-3, 1]$	$[2, 4]$

157 учасників пробного інтернет-тестування так розподілилися за вибором варіантів відповіді на це завдання у порогових групах (таблиця 8.2 та рисунок 8.1):

Таблиця 8.2. Вибір варіантів відповіді членами порогових груп

Порогова група	Варіант				
	А	Б	В	Г	Д*
1	41.94%	6.45%	12.90%	22.58%	16.13%
2	54.84%	6.45%	6.45%	9.68%	22.58%
3	37.50%	3.13%	3.13%	6.25%	50.00%
4	9.68%	0.00%	3.23%	0.00%	87.10%
5	6.45%	0.00%	0.00%	0.00%	93.55%

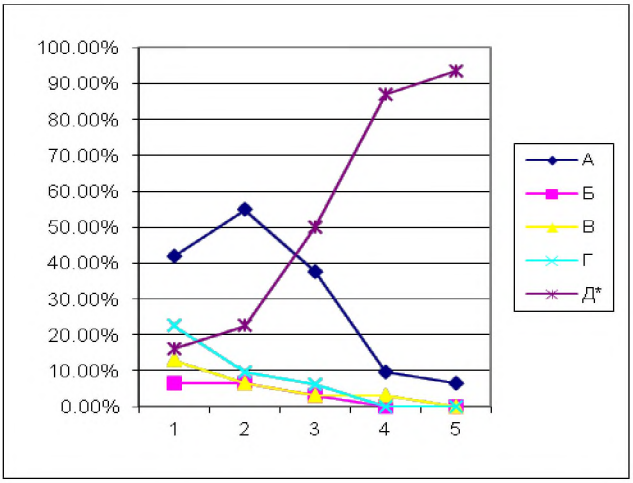

Рис. 8.1. Діаграма розподілу вибору відповідей членами порогових груп

Як видно з таблиці, у двох найслабкіших групах більшість обирали дистрактор **А**: у групі №1 найслабкіших екзаменованих його вибрали 41%, у групі №2 – 54%. Лише у третій групі вибір правильної відповіді **Д** починає переважати, а у найсильнішій групі №5 правильну відповідь обрали 93.55%. Цікаво відмітити, що у цій групі всі інші вибрали все ж варіант **А**. Загалом цей дистрактор вибрали близько 30% від усіх учасників тестування. Враховуючи, що правильну відповідь вибрали приблизно 54% екзаменованих, згідно з правилами побудови завдань множинного вибору, на вибір кожного з чотирьох дистракторів має припадати близько 12%. Чи означає це, що перелік варіантів відповідей на це завдання є недостатньо якісним? І так і ні. Як ми пам'ятаємо, першим і ключовим етапом конструювання тесту є визначення його мети. Для діагностичного тесту це завдання є інформативним, оскільки аналіз дистракторів дозволяє виявити типові помилки у судженнях екзаменованих. Для нормо-орієнтованого тестування високої відповідальності це завдання можливо, підходить менше, оскільки деяка частина сильніших екзаменованих могла б вибрати правильну відповідь, якби не було «провокуючого» дистрактора **А**. Важливо зрозуміти, що рекомендація добиватися приблизно рівних ймовірностей вибору дистракторів є швидше технічною: вона допомагає контролювати їх правдоподібність.

Дискримінативність завдання. Розробника тесту завжди цікавить, наскільки добре дозволяє тестове завдання диференціювати екзаменованих за рівнем вираженості вимірюваної якості. Раніше ми побачили, що завдання з рівнем трудності, близьким до середньої, начебто добре відповідає цій меті. Але уявимо собі таку ситуацію: на деяке завдання прийнятної трудності дають правильну відповідь особи, у яких за іншими завданнями рівень вираженості вимірюваної якості низький, і навпаки, особи з високим рівнем дають неправильну відповідь. Терміни «правильна відповідь» і «неправильна відповідь» вживаються тут у тому сенсі, що саме правильна відповідь повинна свідчити на користь більшої вираженості риси чи конструкту, як це має місце у тестах навчальних досягнень. Отже, для даного завдання ми отримали парадоксальну картину. У цьому випадку кажуть, що завдання має *від'ємну дискримінативність* (або *від'ємну роздільну здатність)*. Таке завдання не може вважатися валідним для даного вимірюван-

ня, і його потрібно вилучити з тесту або знайти у ньому помилку. Але й для цілком валідних завдань однакової трудності можна спостерігати різну здатність диференціювати опитуваних.

Існують різні показники дискримінативності. Один з них заснований на методі порогових оцінок (порогових груп), інші – на понятті кореляції. Розглянемо деякі з найбільш уживаних показників дискримінативності.

Індекс дискримінативності (чи *індекс роздільної здатності*) – найпростіший та найчастіше вживаний показник. Його використовують для дихотомічних завдань. Як і у випадку аналізу дистракторів методом порогових груп, для обчислення індексу потрібна інформація про результати тестування повним тестом або будь-який інший, зовнішній, критерій, за яким з групи екзаменованих, які брали участь у апробації завдання, відбирають дві підгрупи. Це можуть бути як половини, так і менші частини групи учасників. Існує дослідження Келлі, яке показує, що за певних широких умов чутливий та водночас стійкий індекс роздільної здатності можна отримати, якщо відібрати до групи найбільш слабких 27% учасників, і стільки ж – до групи найбільш сильних. Нехай p_u, p_l – частки учасників, які відповіли на завдання правильно, відповідно у групі найсильніших та групі найбільш слабких. Тоді індекс дискримінативності обчислюється за формулою:

$$D = p_u - p_l.$$

Нехай, наприклад, з групи 100 учасників тестування вибрано 27 (27%) найбільш слабких за загальним результатом, та 27 найбільш сильних. Якщо у групі сильних на дане завдання відповіли правильно 18 екзаменованих, а у групі слабких – 3 екзаменованих, то частки становитимуть: $p_u = \frac{18}{27} = \frac{2}{3}, p_l = \frac{3}{27} = \frac{1}{9}$, і індекс дискримінативності для цього завдання становитиме

$$D = \frac{6}{9} - \frac{1}{9} = \frac{5}{9} \approx 0{,}56.$$

Оскільки максимально можливе значення частки становить 1 (всі представники підгрупи відповіли правильно), а мінімально можливе – 0 (жоден не відповів правильно), то значення індексу

дискримінативності завжди лежить у межах від –1 до +1. Від'ємні та близькі до нуля значення вказують на погану якість завдання, його потрібно переробити або вилучити з тесту. Вважається, що завдання задовільно диференціює екзаменованих за рівнем вираженості вимірюваної якості, якщо $D \geq 0{,}4$.

Оскільки рішення про те, який відсоток осіб повинен потрапити до порогових груп, розробник тесту приймає самостійно, інформацію про це потрібно давати у супровідній документації до тесту. Простота цього показника таїть у собі його недоліки. Зокрема, розподіл індексу роздільної здатності невідомий, і це не дозволяє визначати математично, чи є, скажімо, відмінність знайденого значення від нуля, або різниця між індексами двох завдань значущими. Тим не менше через легкість обчислення та інтерпретації індекс дискримінативності залишається найбільш уживаним показником роздільної здатності завдання, особливо у тестуваннях на рівні класу чи студентської групи.

У випадках, коли завдання оцінюється за політомічною шкалою (якою є, наприклад, популярна шкал Лайкерта, що використовується у психологічних тестах), для визначення роздільної здатності використовують коефіцієнт кореляції. Обчислюється кореляція між оцінками учасників апробації тесту за дане завдання та їх оцінками за весь тест або за зовнішній критерій. Далі розглянемо чотири варіанти обчислення коефіцієнта кореляції:

- точково-бісеріальна кореляція;
- бісеріаальний коефіцієнт кореляції;
- фі-коефіцієнт;
- тетрахоричний коефіцієнт кореляції.

Коефіцієнт *точково-бісеріальної* (*point-biserial*) *кореляції* обчислюється для дихотомічного завдання. Його вибіркова формула має вигляд:

$$\rho_{pbis} = \frac{\bar{X}_+ - \bar{X}}{s_X}\sqrt{p/q},$$

де $\bar{X}_+$ – середня критеріальна оцінка (тобто оцінка за тест або зовнішній неперервно розподілений критерій) тих учасників, які від-

повіли на дане завдання правильно, $\bar{X}$ – середня критеріальна оцінка для всіх учасників, s_X – вибіркове стандартне відхилення критеріальної оцінки, p – трудність завдання, $q = 1 - p$.

Якщо критеріальною оцінкою є оцінка за тест, до якого входить дане завдання, то значення коефіцєнта кореляції є дещо завищеним через те, що оцінки за дане завдання входять і до загальної оцінки. Якщо кількість завдань у тесті достатньо велика (більше 25), то це не створює проблеми. Якщо ж завдань мало, то можна скористатися формулою:

$$\rho_{i(X-i)} = \frac{\rho_{Xi}s_X - s_i}{\sqrt{s_i^2 + s_X^2 - 2\rho_{Xi}s_X s_i}},$$

де $\rho_{i(X-i)}$ – коефіцієнт кореляції між завданням та тестом, з якого це завдання видалене.

Іншим показником дискримінативності дихотомічного завдання є *бісеріальний* коефіцієнт кореляції. В основі використання цього показника лежить припущення, що вимірювана якість розподілена у цільовій популяції за нормальним законом.

Вибіркова формула для бісеріального коефіцієнта:

$$\rho_{bis} = \frac{\bar{X}_+ - \bar{X}}{s_X}\sqrt{p/Y},$$

де усі позначення, крім Y, мають той же зміст, що й у формулі для точково-бісеріального коефіцієнта, а Y – це ордината кривої щільності стандартного нормального розподілу у точці з абсцисою, що дорівнює z-оцінці, яка відповідає трудності завдання p. Наприклад, для завдання з трудністю 0,6 z-оцінка дорівнює 0,25, а відповідна ордината нормальної кривої дорівнює 0,3867. Знайти це значення можна за спеціальною таблицею значень стандартного нормального розподілу, пам'ятаючи, що величина p – це ймовірність, тобто площа під кривою щільності, обмежена справа прямою $y = z$.

Слід пам'ятати, що точково-бісеріальний та бісеріальний коефіцієнти – це різні величини, які не збігаються. Математично зв'язок між ними виражається формулою

$$\rho_{bis} = \frac{\sqrt{pq}}{Y}\rho_{pbis}.$$

Знаменник дробу у цій формулі завжди менший від чисельника, тобто значення бісеріального коефіцієнта завжди є більшим за відповідне значення точково-бісеріального коефіцієнта. Різниця становить мінімум 1,5 разів. Для завдань помірної трудності ця різниця є меншою, ніж для дуже легких чи дуже складних завдань. На кінцях розподілу трудності ця різниця може досягати 4 разів. Таким чином, розробники тесту зобов'язані у супровідній документації вказувати, який саме варіант коефіцієнта кореляції обчислювався для визначення дискримінативності завдань.

Буває, що критерій, з яким потрібно порівняти оцінки за дихотомічне завдання, сам є дихотомічним (наприклад, стать екзаменованого, отримання чи не отримання ним заліку тощо). У таких випадках як показник дискримінативності завдання може використовуватися фі-коефіцієнт кореляції, який є формою коефіцієнта кореляції Пірсона для дихотомічних змінних. Цю величину ми ввели у главі 2. Нагадаємо формулу для фі-коефіцієнта кореляції:

$$\rho_{\varphi} = \frac{p_{jk} - p_j p_k}{\sqrt{p_j q_j p_k q_k}},$$

У нашому випадку один із індексів при величинах трудності і «легкості» позначає досліджуване завдання, інший індекс позначає критерій.

Зауважимо, що використання фі-коефіцієнта є виправданим лише тоді, коли дихотомічність критерію є природною, а не створюється штучно заради спрощення обчислень, оскільки при спрощенні істотно втрачається чутливість показника, тобто здатність його розрізняти різні завдання за дискримінативністю. Значення 1 цей коефіцієнт досягає лише в одному випадку – коли обидві змінні – завдання і критерій мають однакову трудність. Як і точково-бісеріальний, цей коефіцієнт є все тим же коефіцієнтом лінійної кореляції Пірсона, а відмінність у назвах коефіцієнтів пов'язана лише із формулами для обчислення.

В окремих випадках дослідник вдається до дихотомізації нормального розподілу. Отримані таким чином дихотомічні змінні можуть досліджуватися на кореляваність за допомогою так званого *тетрахоричного* коефіцієнта. Цей показник вільний від недоліку фі-коефіцієнта у досяганні значення 1, тобто у тих випадках, коли частки осіб, які справилися з завданням, і які справилися з критерієм, є різними. Оскільки цей показник використовується дуже рідко, а його формула є достатньо громіздкою, не будемо її наводити.

Ми розглянули п'ять різних показників роздільної здатності тестового завдання. Який з них для яких випадків найкраще підходить? При виборі показника дискримінативності можна керуватися наступними правилами.

1. Якщо завдання мають близьку до середньої трудність, то вибір того чи іншого показника не має особливого значення. При цьому використання індексу D є найпростішим, але коли вимагається перевірити знайдене значення показника на значущість, потрібно обирати один з коефіцієнтів, заснованих на кореляції.

2. Якщо досліджується завдання екстремальної трудності, то, за умови, що розподіл вимірюваної якості у цільовій популяції близький до нормального, краще використовувати бісеріальний коефіцієнт.

3. Якщо дослідник не впевнений у тому, що майбутні вибірки осіб не будуть сильно відрізнятися від даної вибірки за трудністю для них даного завдання, краще використовувати бісеріальну кореляцію, оскільки мале значення цього коефіцієнта для вибірок опитуваних з загалом високим або низьким рівнем вимірюваної якості свідчить саме про низьку роздільну здатність завдання, а не є просто функцією від трудності завдання, як це може бути для цього випадку з точково-бісеріальним коефіцієнтом.

4. Якщо розробник тесту передбачає, що майбутні вибірки опитуваних мало відрізнятимуться від даної за рівнем вимірюваної якості, а метою є вибір таких завдань, які забезпечують високу внутрішню узгодженість тесту, то існують підстави для вибору точково-бісеріального коефіцієнта.

5. Якщо і завдання, і критерій є дихотомічними, можна використовувати фі-коефіцієнт або тетрахоричний коефіцієнт. При цьому більш складний для обчислення тетрахоричний коефіцієнт

використовується коли дихотомічні змінні отримані штучно з нормально розподілених величин, частки тих, хто справився з завданням і тих, хто справився з критерієм, істотно відрізняються між собою, і знайдені кореляції планується використовувати у факторному аналізі.

Показники надійності та валідності завдань. Хоча розглянуті вище показники трудності та дискримінативності завдань можуть допомагати в комплексному дослідженні тесту на валідність та надійність, існують показники, які більш безпосередньо пов'язані з цими поняттями. Вони є одночасно функціями як від корельованості завдання з критерієм, так і від мінливості оцінок за завдання.

Показником надійності i-го завдання називається величина $s_i\rho_{iX}$, де ρ_{iX} – коефіцієнт кореляції між оцінкою за завдання та критерій. Для дихотомічного завдання її можна записати як $\sqrt{p_i q_i}\rho_{iX}$, де ρ_{iX} – коефіцієнт точково-бісеріальної кореляції між оцінкою за завданням та оцінкою за весь тест. Дисперсія по завданню є фактично вагою внеску завдання у загальну надійність тесту, тому при потребі отримати тест з високою надійністю при відборі завдань до нього слід замість простих коефіцієнтів кореляції контролювати показники надійності завдань.

Також можна показати, що дисперсія загальної тестової оцінки дорівнює квадрату суми показників надійності завдань:

$$s_X^2 = \left(\sum s_i\rho_{iX}\right)^2.$$

Цим фактом зручно користуватися, підбираючи завдання до тесту з заданим рівнем дисперсії. Так само, якщо розробником покладений мінімум коефіцієнта альфа внутрішньої узгодженості тесту, він може контролювати зміну цього показника з додаванням до тесту кожного нового завдання, користуючись формулою

$$\rho_\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{(\sum s_i\rho_{iX})^2}\right),$$

де k – кількість завдань тесту на даний момент.

Нарешті, якщо розробник, щоразу додаючи до тесту нове завдання, хоче контролювати коефіцієнт валідності у вигляді кореляції між створеним набором завдань та зовнішнім критерієм Y, він може це робити за допомогою формули

$$\rho_{XY} = \frac{\sum s_i \rho_{iY}}{\sum s_i \rho_{iX}}.$$

Розмір вибірки. Загального правила для планування мінімального розміру вибірки осіб з цільової популяції для апробації тестових завдань не існує. Очевидно, чим більшою є вибірка, тим надійнішими будуть отримані на її основі оцінки параметрів тестових завдань. Для апробації широкомасштабних тестів регіонального чи національного рівня бажано мати вибірку об'єму не менше 200 осіб (сказане стосується дослідження параметрів, описаних раніше. Для оцінки параметрів на основі теорії IRT, про яку йтиметься пізніше, потрібна більша вибірка). Існує також емпіричне правило, згідно з яким кількість осіб у вибірці повинна перевищувати кількість завдань у тесті мінімум у 5 разів.

Тактика відбору завдань до тесту. При відборі до тесту вже апробованих завдань виникає одна з двох загальних ситуацій. У першій ситуації завдань є набагато більше, ніж передбачається використовувати у тесті. Практично завжди розробник намагається отримати тест заданої якості з якомога меншим набором завдань, оскільки це економить час та інші ресурси. Тому потрібно поступово додавати до тесту ті завдання, які дають найбільший внесок у бажаний рівень надійності та валідності тесту. Вище ми розглянули показники надійності та валідності завдань та методи контролю на їх основі надійності та валідності тесту в цілому.

В іншій ситуації розробник не володіє надто великою сукупністю тестових завдань, і тому намагається зберегти кожне завдання, вклад якого у контрольовані параметри є позитивним. Спочатку можна залишати у тесті всі завдання, які забезпечують достатньо високу кореляцію з критерієм.

Важливим є контроль стандартної похибки для коефіцієнта кореляції. Якщо дискримінативність завдання оцінювалося за допомогою коефіцієнта точково-бісеріальної кореляції або коефіцієн-

нта фі, можна скористатися зручною наближеною формулою: $s_\rho = 1/\sqrt{N-1}$, де N – об'єм вибірки. Цю формулу можна використовувати при $N \geq 50$. Наприклад, якщо у вибірці 101 особа, то за цією формулою $s_\rho = 0{,}1$. Зазвичай мінімальне критичне значення покладається на 2 стандартні похибки вище нуля. Для нашого прикладу це буде 0,2. Отже, потрібно залишити у тесті ті завдання, значення точково-бісеріального коефіцієнта яких не менше за 0,2.

Якщо для оцінки роздільної здатності завдань використовувався бісеріальний коефіцієнт кореляції, то стандартну похибку можна оцінити за формулою

$$s_{bis} = \frac{\sqrt{pq/N-1}}{Y},$$

де зміст позначень у правій частині – той же, що й у формулі для бісеріального коефіцієнта кореляції. Потрібно пам'ятати, що стандартна похибка бісеріальної кореляції мінімальна для завдань середньої трудності, і зростає із наближенням трудності до мінімальної та максимальної.

Як повинні впливати на відбір завдань дані про їх трудність? Для нормо-орієнтованого тестування первинним показником завдання є не трудність, а дискримінативність. Трудність завдання може істотно відрізнятися для різних вибірок. Похибку вибірки в оцінці трудності можна оцінити за формулою $s_p = \sqrt{pq/N}$. Для тесту, який, очікувано, буде надійно диференціювати осіб цільової популяції по широкому діапазону вимірюваної якості, завдання повинні мати середню трудність у діапазоні 0,4-0,6. Раніше ми також зауважили, що для кращої диференціації на кінцях цього діапазону бажано, щоб до тесту входила також мала кількість завдань високої та низької трудності. Особливо це стосується випадку, коли середня бісеріальна кореляція між завданнями та загальною тестовою оцінкою перевищує 0,6. Також включення до тесту завдань з екстремальним рівнем трудності потрібне у випадку, коли тестування передбачає прийняття рішення щодо осіб з екстремальним рівнем вираженості вимірюваної якості, наприклад, при відборі малої кількості осіб з великої кількості кандидатів для відповідальної роботи.