Ю.О. Ковальчук

Теорія освітніх вимірювань

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

Книгу видано в рамках міжнародного проекту «Освітні вимірювання, адаптовані до стандартів ЄС»
за програмою Європейського Союзу Темпус

Ковальчук Ю.О.
Теорія освітніх вимірювань. – Ніжин: Видавець ПП Лисенко М.М., 2012. – 200 с.

Книга містить виклад основ теорії освітніх вимірювань на матеріалі лекцій, які читаються автором студентам магістратури зі спеціальності «Освітні вимірювання» в Ніжинському державному університеті імені Миколи Гоголя. Книга може бути корисною студентам та аспірантам педагогічних спеціальностей, викладачам та іншим працівникам освітньої галузі, працівникам з найму та управління трудовими ресурсами.

ISBN **ББК 74.04(4Укр)я73**

© Ковальчук Ю.О., 2012

ЗМІСТ

ПЕРЕДМОВА

У 1993 році в Україні була здійснена спроба запровадити тестування випускників загальноосвітніх шкіл, однак ця спроба виявилася невдалою. Лише з 2002 року розпочалося успішне становлення вітчизняної системи зовнішнього незалежного оцінювання. Однією з головних причин перших невдач упровадження в український освітній простір методів об'єктивного оцінювання навчальних досягнень учнів, абітурієнтів та студентів є тривале ігнорування радянською педагогічною наукою теорії і практики освітніх вимірювань як специфічного розділу психометрії.

Починаючи з 2008 року, завдяки зусиллям Міністерства освіти і науки, Академії педагогічних наук України, та за підтримки програми Європейського Союзу Tempus у трьох українських університетах впроваджено програми підготовки фахівців з освітніх вимірювань бакалаврського та магістерського рівнів. Тим самим розпочато процес ліквідації відриву вітчизняної педагогічної науки від більш ніж столітньої світової традиції освітніх вимірювань.

Ця книга містить виклад основних розділів теорії освітніх вимірювань на тому рівні й приблизно у тому об'ємі, який відповідає навчальному курсу з освітніх вимірювань, який вивчається студентами-магістрантами у Ніжинському державному університеті імені Миколи Гоголя.

Автор є прихильником широкого тлумачення освітніх вимірювань як розділу психометрії, який охоплює не лише область тестування рівня навчальних досягнень, а й вимірювання усіх інших якостей осіб, що навчаються, якщо результати цих вимірювань можуть становити інтерес для педагогів, фахівців з найму та управління трудовими ресурсами, психологів – усіх тих, кому так чи інакше доводиться у своїй професійній діяльності вирішувати питання підвищення ефективності навчання. Вимірювання як вищий прояв оцінювання повинне зайняти чільне місце в системі моніторингу якості освіти як у навчальних закладах усіх рівнів, так і на робочих місцях, а також допомагати тим, хто займається самоосвітою.

Оскільки єдиним на сьогодні інструментом вимірювання можна вважати тестування, у книзі ці терміни вживаються як синоніми. Автор свідомо не дає визначення тесту, щоб уникнути неминучого в такому випадку штучного звуження цього складного та багатогранного поняття.

Через обмежений об'єм книги до неї не увійшли такі важливі розділи, як історія становлення психометрії та освітніх вимірювань у світі, забезпечення справедливості тестування, особливості критеріально-орієнтованого вимірювання.

Логіка викладу матеріалу в основному відповідає класичному підручнику Лінди Крокер та Джеймса Алгіни [6], звідти ж почерпнуто ряд ілюстративних прикладів. В основу деяких глав, передусім глави 4, присвяченої загальному огляду поняття валідності вимірювань, глав 10 (шкалювання) та 11 (порівняння результатів вимірювань), покладено матеріали 4-го видання фундаментальної колективної праці Educational Measurement [10]. Використовувалися й інші джерела, серед яких можна виділити книги М. Б. Челишкової [9] та А. Анастазі й С. Урбіни [1]. В кінці книги наведено список рекомендованих для самостійного опрацювання джерел, який свідомо обмежено автором за критеріями якості і доступності.

Розуміючи, що успішне опанування матеріалом книги вимагає певної математичної підготовки читачів, і бажаючи, щоб книга могла бути корисною якомога ширшому їх колу, автор присвятив главу 2 базовим поняттям математичної статистики у розрізі їх застосування до освітніх вимірювань. У тих місцях, де вимагаються більш ґрунтовні знання спеціальних математико-статистичних методів, таких як факторний чи дискримінантний аналіз, читачу пропонується звернутися до спеціальної літератури.

1. ПОНЯТТЯ ОСВІТНЬОГО ВИМІРЮВАННЯ

Три теорії вимірювань. Людина, яка вперше зустрічається з ідеєю вимірювання таких феноменів, як «інтелект», «навченість», «здібність», відчуває збентеження. У повсякденному житті вимірювання застосовуються до фізичних властивостей предметів, таких як довжина, вага, об'єм, тощо. Фізичне вимірювання має ряд властивостей, які важко перенести на сферу психічного. Важливою властивістю фізичного вимірювання є адитивність, або існування одиниці вимірювання такої, що міра предмета є простою сумою цих одиниць, так, якби ми при вимірюванні відкладали послідовно якийсь «твердий брусок», рахуючи кількість відкладень.

Таке розуміння вимірювання існувало у фізиці століттями, аж до тих часів, коли фізична наука проникла у мікросвіт. Розвиваючи квантову механіку, фізики зіштовхнулися з питаннями: 1) що ми вимірюємо? 2) Як вимірювальний прилад може впливати на сам процес вимірювання?

Для психології та інших суспільних наук ці питання були актуальними від самого початку. Якщо у фізиці дослідник взаємодіє з суб'єктом вимірювання через вимірювальний прилад, то в суспільних науках таким «приладом» є вся організація експерименту. Потреба суспільних наук у вимірюваннях змусила дослідників переосмислювати саме поняття вимірювання. У 19-20 століттях спостерігались три підходи, або три теорії вимірювань, які використовувались у психології: класична, операціональна та репрезентативна.

У відповідності з *класичною теорією вимірювань*, психологія є кількісною наукою. Засновник психофізіології Густав Фехнер вважав, що в загальному випадку вимірювання кількості полягає у з'ясуванні, як часто одиниця кількості одного і того ж виду міститься у ній. У наш час класична теорія практично повністю витіснена репрезентативною та операціональною теоріями.

Згідно з *операціональною теорією вимірювань*, яка розвивалася у першій половині 20-го століття, під будь-яким концептом (тобто поняттям) розуміється не що інше як сукупність операцій. З операціональної точки зору будь-яке вимірювання – це операція,

що породжує числа. Числа є самодостатніми і не залежать від природи об'єкта, щодо якого відбувається вимірювання. В граничному розумінні для операціоналіста наука є просто вивченням виконуваних операцій, а не дослідженням реальності. Так, операціоналіст розглядає результати тестування як вимірювання вже тоді, коли вони є наслідком осмисленого приписування чисел, виконаного за допомогою строго визначених операцій.

Для послідовника *репрезентативної теорії вимірювань*, отриманих в результаті тестування оцінок недостатньо для того, щоб стверджувати, що вони є результатом вимірювання, хоча й отримані внаслідок чітко визначених операцій. Для того, щоб оцінки можна було вважати результатом вимірювань, необхідно, щоб відношення між ними (скажімо, відношення порядку) репрезентували якісні емпіричні відношення між успішністю виконання тесту різними особами, і дослідження мають полягати в ідентифікації таких відношень і описанні їх характеристик.

Останнім часом в психологічних і освітніх вимірюваннях більш широко використовуються положення саме репрезентативної теорії вимірювань.

Репрезентативна теорія вимірювань була розроблена американським психологом Стенлі Стівенсом в 40-х роках 20 століття. Центральним поняттям цієї теорії є поняття *вимірювальної шкали*.

Слід розрізняти *об'єкти вимірювання*, їх *властивості* і *ознаки*. Останні виражають у тій чи іншій числовій шкалі властивості об'єктів.

В залежності від того, яка операція лежить в основі вимірювання ознаки, виділяють *шкали вимірювання*, які часто називають ще *рівнями вимірювання*. Тип шкали задає групу *допустимих перетворень* шкали. Допустимі перетворення не міняють співвідношень між об'єктами вимірювання, тобто після допустимих перетворень дані залишаються виміряними у шкалі того ж самого типу. Наприклад, при вимірюванні довжини перехід від футів до метрів не міняє співвідношень між довжинами даних об'єктів – якщо перший об'єкт довший другого, то це буде встановлено і при вимірюванні у футах, і при вимірюванні в метрах. При цьому чисельне значення довжини в футах відрізняється від чисельного значення довжини в метрах – не міняється лише результат порівняння довжин двох об'єктів.

Перш за все, шкали поділяються на метричні і неметричні (за іншою термінологією, шкали якісних ознак і шкали кількісних ознак, або шкали низького рівня і шкали високого рівня).

Шкала є *метричною*, якщо для неї може існувати одиниця вимірювання. Також шкали поділяються на *дискретні* та *неперервні*.

Типи шкал. Розглянемо основні чотири типи шкал.

1. Номінальна шкала (шкала найменувань, категоріальна шкала). У шкалі найменувань допустимими є тільки взаємнооднозначні (тотожні) перетворення. Використання цієї шкали має на меті класифікацію об'єктів. В цій шкалі числа якщо й використовуються, то лише як мітки для розрізнення об'єктів. В шкалі найменувань виміряні, наприклад, номери телефонів, автомашин, паспортів, студентських квитків. Номери страхових свідоцтв державного пенсійного страхування, медичного страхування, штрихкоди товарів теж виміряні в шкалі найменувань. Стать людей також вимірюється в шкалі найменувань, результат вимірювання приймає два значення – чоловіча, жіноча. Раса, національність, колір очей, волосся – номінальні ознаки. Нікому не прийде в голову додавати або множити номери телефонів, такі операції не мають для номінальної шкали сенсу. Єдине, для чого годяться результати вимірювань в шкалі найменувань – для розрізнення об'єктів. У багатьох випадках цього достатньо. Наприклад, шафки для одягу в роздягальнях для дорослих розрізняють за номерами, тобто числами, а в дитячих садках використовують малюнки, оскільки діти ще не знають чисел.

2. Порядкова (рангова) шкала. Вище ми розглянули приклад з шафками для одягу у дорослій і дитячій роздягальнях. У тому випадку, коли ми вирішимо позначати шафки не малюнками, а числами, ми можемо закласти в ці позначення додаткову інформацію, яка буде допомагати швидше відшукати потрібну шафку. А саме, ми можемо впорядкувати розташування шафок за зростанням чи спаданням чисел, написаних на них (іншими словами, пронумеруємо шафки). Очевидно, зменшення невизначеності дозволяє стверджувати, що ми переходимо таким чином до шкали вищого типу.

У порядковій шкалі числа використовуються не тільки для розрізнення об'єктів, але і для встановлення порядку між об'єкта-

ми. Простим прикладом є оцінки знань. Символічно, що оцінки студентів за чотирибальною національною шкалою можуть виражатись як числами (2, 3, 4, 5), так і словесно – незадовільно, задовільно, добре, відмінно. Цим підкреслюється "нечисловий" характер оцінок. Не можна стверджувати, що учень, який отримав двійку, знає рівно вдвічі менше, ніж той, що отримав четвірку, або різниця між тим, хто отримав трійку і тим, хто отримав четвірку, дорівнює різниці у знаннях між тими, хто отримав відповідно четвірку і п'ятірку.

Оцінки експертів найчастіше слід вважати виміряними в порядковій шкалі, адже людина більш правильно (і з меншими утрудненнями) відповідає на питання якісного, наприклад, порівняльного, характеру, ніж кількісного. Так, наприклад, легше сказати, який з двох відрізків довший, ніж вказати їх довжину в міліметрах.

У різних областях людської діяльності застосовується багато видів порядкових шкал. Так, наприклад, в мінералогії використовується шкала Мооса, по якому мінерали класифікуються згідно критерію твердості. Так, згідно з цією шкалою, тальк має бал 1, гіпс – 2, а алмаз – 10. Мінерал з більшим номером є більш твердим, ніж мінерал з меншим номером, при натисненні дряпає його.

Порядковими шкалами в географії є – бофортова шкала вітрів («штиль», «слабкий вітер», «помірний вітер» і т.д.), шкала сили землетрусів. Очевидно, не можна стверджувати, що землетрус в 2 бали (лампа гойднулася під стелею – таке буває і в нашій місцевості) рівно в 5 разів слабкіший, ніж землетрус в 10 балів (повне руйнування всіх споруд на поверхні землі).

У медицині є багато загальноприйнятих порядкових шкал – шкала стадій гіпертонічної хвороби, шкала опіків тощо. Номери будинків також виміряні в порядковій шкалі – вони показують, в якому порядку стоять будинки уздовж вулиці. Номери томів в зібранні творів письменника або номери справ в архіві підприємства зазвичай пов'язані з хронологічним порядком їх створення, тому теж є прикладами порядкової шкали.

До порядкових шкал *допустимими є перетворення, які не порушують порядку між результатами вимірювань, тобто будь-які монотонні перетворення*, наприклад, логарифмування. Слід пам'ятати, що спадне перетворення змінює порядок розміщення чисел на шкалі на протилежний.

При статистичному опрацюванні порядкових даних відшукання такої характеристики, як середнє арифметичне, взагалі кажучи, не має сенсу.

Порядкова шкала і шкала найменувань є основними шкалами якісних ознак і відносяться до *неметричних шкал.* В багатьох конкретних областях науки і практики результати якісного аналізу можна розглядати як вимірювання за цими шкалами.

Шкали кількісних ознак (метричні шкали, або шкали вищого типу) – це, в основному, шкала інтервалів і шкала відношень.

3. Шкала інтервалів. У цій шкалі числа відображають не тільки відмінність між об'єктами за ступенем вираженості ознаки, а й *на скільки* більше або менше виражена ознака. Характерною особливістю цієї шкали, яка відрізняє її від іншої метричної шкали – шкали відношень, – є *відсутність природного початку відліку і природної одиниці вимірювання*. Дослідник повинен сам задати точку (початок) відліку і сам вибрати одиницю вимірювання.

Допустимими перетвореннями в шкалі інтервалів є будь-які лінійні перетворення, тобто перетворення виду $y = ax + b$. Наприклад, температурні шкали Цельсія і Фаренгейта зв'язані залежністю $°C = 5/9\,(°F - 32)$, де °C – температура (в градусах) за шкалою Цельсія, а ^{0}F – температура за шкалою Фаренгейта.

Для інтервальної шкали величина середнього арифметичного має цілком визначений зміст.

4. Шкала відношень. З кількісних шкал найпоширенішими в природничих науках і у повсякденній практиці є шкали відношень. У них є *природний початок відліку* – нуль, який означає відсутність величини. В такій шкалі числа відображають не тільки відмінність між об'єктами за ступенем вираженості ознаки, не тільки те, на скільки більше або менше, а й *у скільки разів* більше або менше виражена ознака.

За шкалою відношень вимірюється більшість фізичних одиниць: маса тіла, довжина, заряд, а також ціни (і різні вартісні характеристики) в економіці. Тільки для інтервальної шкали результати вимірювань є числами в звичному значенні слова. Допустимим перетворенням для шкали відношень є *масштабування* (розтягування або стискання) виду $y = ax$.

Слід зазначити, що за іншою, більш детальною класифікацією, шкалу відношень відрізняють від так званої *абсолютної* шка-

ли. Відмінність цих шкал полягає в тому, що *для шкали відношень не існує природної одиниці вимірювання, а для абсолютної шкали така одиниця існує*. Прикладом віднесення до абсолютної шкали з природною одиницею вимірювання є кількість людей в кімнаті.

Перераховані шкали можна характеризувати за їх *диференційовною (роздільною) здатністю (чи потужністю)* – здатністю розрізняти об'єкти як відмінні один від одного.

Шкали по мірі зростання потужності розташовуються так: *шкала найменувань, порядкова, інтервальна, відношень*. Тобто неметричні шкали менш потужні – вони несуть менше інформації про відмінності між об'єктами.

Шість рівнів наукового усвідомлення. У процесі розвитку відповідної області знання тип шкали може мінятися. Так, спочатку температура вимірювалася за порядковою шкалою (холодніше – тепліше). Потім – за інтервальною (шкали Цельсія, Фаренгейта, Реомюра). Нарешті, після відкриття абсолютного нуля температуру можна вважати виміряною за шкалою відношень (шкала Кельвіна). Зауважимо, що серед фахівців іноді виникають розбіжності з приводу того, за якими шкалами слід вважати виміряними ті або інші реальні величини. Допомогти тут можуть запитання на зразок такого: «якщо сьогодні температура 0 градусів, а завтра буде удвічі холодніше, то якою буде завтра температура?».

Можна прослідкувати зв'язок між рівнями вимірювання і рівнями науковості мислення, слідуючи запропонованій Ч. Пірсом класифікації, яку він розглядав в рамках створеної ним семіотики – науки про знаки. Ця класифікація відображає еволюцію людського мислення – від перших «диких» спалахів уяви до репродуктивних кількісних законів, які є інструментами і вищими проявами науки.

Власне, класифікація знаків у Пірса налічує десять рівнів, але для нас є суттєвими вищі шість щаблів. Охарактеризуємо коротко кожен з них.

1. Уявлення. Перший рівень наукового усвідомлення – це просте інтуїтивне уявлення, спалах думки, «дика» гіпотеза. Це «зерна» творчості. Таке уявлення належить тільки даній особі, не може нею бути переосмислене чи передане комусь іншому. Пірс називає цей перший рівень обізнаності «можливою іконою» (possible icon).

2. Думка. Деякі уявлення вимагають осмислення. З'являються думки, які тепер можуть відтворюватися у нашому мозку неодноразово. Ми також можемо тепер ідентифікувати ті ж самі думки у інших осіб. Думка, таким чином, стає тим, до чого ми можемо повертатися і чим ми можемо обмінюватися з іншими людьми. Це цілком якісна характеристика. Eddington (1946) та Kinston (1985) називають цей рівень «суттю» (entity). Пірс називає це «можливим індексом» (possible index).

3. Об'єкт. Наступний крок – усвідомлення того, як ми можемо вказати на нашу думку. Ми присвоюємо нашій думці ім'я, і тепер ми можемо поводитися з нею, як з річчю, частиною світу, екземпляром нашої ідеї. Ми тепер можемо вказувати на ідею, і можемо підраховувати ідеї. На цьому рівні з'являється поняття кількості. Це той рівень, на якому може існувати перший, за Стівенсом, рівень (шкала) вимірювань, який називається номінальним чи категоріальним. На цьому рівні ми можемо віднести даний об'єкт до певного класу, наприклад, віднести особу до класу чоловіків чи до класу жінок. Клас може позначатися числом, але це тількм символ, який не має сенсу числа. Eddington (1946) та Kinston (1985) називають цей рівень «спостережуваним» (observable). Пірс називає це «дійсним індексом» (factual index).

4. Порівняння. Далі приходимо до ідеї надання переваги. Деякі речі можуть подобатися нам, інші – ні. Ми починаємо враховувати «кращість» чи «гіршість», присвоюючи їм відповідні числа.

Хоча ми ще не кажемо, на скільки щось краще за інше, чи більше за інше, чи важче за інше. Стівенс назвав це порядковим рівнем вимірювання. За Пірсом, це «можливий символ» (possible symol). Саме тут починає з'являтися ідея подрібнюваності, квантування. Ми переходимо від підрахунку «солодких яблук» до ідеї «солодкості».

5. Вимірювання. Підрахунок – це початок вимірювань. Однак проста кількість не завжди приводить до того, що ми називаємо мірою. Якщо цеглини мають різні розміри, то ми не можемо сказати, скільки цеглин нам потрібно, щоб збудувати задумане. Три яблука не завжди будуть оцінені у ту ж саму ціну. На цьому рівні ми приходимо до вигадування абстрактної змінної, вздовж якої ми можемо відкладати рівні частки, чи інтервали.

Стівенс назвав цей рівень інтервальним вимірюванням. Пірс назвав це «дійсним символом» (factual symbol).

6. Залежність. Шостий, найвищий рівень – це рівень виникнення теорії. Що дає нам якась змінна, якщо вона не веде до іншої? Як змінні описують процес? Якби не існувало цього рівня наукового усвідомлення, то не існувало б і поняття статистичного аналізу даних вимірювання, не існувало би дослідження мінливості, кореляції, не існувало би поняття моделі і порівняння моделей.

Пірс називає це «символом, який вимагає доведення» (arguable symbol), а Кінстон – «пов'язуваним».

Описане співвідношення рівнів вимірювання з рівнями наукового усвідомлення свідчать про те, що вимірювання у якійсь обраній шкалі – це не просто «окремий спосіб» вимірювання. Кожен наступний рівень вимірювання ґрунтується на попередніх, включає їх у себе, є їх подальшим розвитком і вдосконаленням.

Формалізація основних понять репрезентативної теорії вимірювань. Назвемо *емпіричною системою з відношеннями* сукупність реальних (емпіричних) об'єктів, які цікавлять дослідника, з визначеною для них системою відношень. Прикладом емпіричної системи з відношеннями може бути група учнів певного класу, які є «носіями» знань з певного предмету, скажімо, фізики, причому для учнів визначені бінарні відношення виду «учень Б знає фізику краще, ніж учень А». Для кожної пари учнів питання про те, хто з них володіє більшими знаннями з фізики, має сенс, тобто є осмисленим. Емпірична система формується дослідником довільно, у відповідності з його уявленнями про реальність, яка вивчається.

Назвемо *математичною системою з відношеннями* сукупність математичних об'єктів (часто такими об'єктами є числа), з виділеними відношеннями між ними. Вимірювання полягає в приписуванні об'єктам емпіричної системи з відношеннями символів математичної системи з відношеннями за допомогою певних правил.

Приписування чисел об'єктам створює числову шкалу. Створення шкали є можливим, оскільки існує ізоморфізм емпіричних систем і систем дій, виконаних над реальними об'єктами. Розрізняють декілька типів числових систем з відношеннями і відповідно кілька типів шкал. Операції, а саме – способи вимірювання об'єктів, задають тип шкали. Шкала, в свою чергу характеризуєть-

ся видом перетворень, які можуть бути віднесені до результатів вимірювання. Якщо не дотримуватися цього правила, то структура шкали порушиться, а дані вимірювання не можна буде осмислено інтерпретувати. Тип шкали також однозначно визначає сукупність статистичних методів, які можуть бути застосовані для обробки даних вимірювання.

Шкала (лат. scala – сходи) у буквальному значенні є вимірювальний інструмент. Нехай A – емпірична система з відношеннями, R – числова система з відношеннями, f – функція, яка гомоморфно відображає A в підсистему R. Зазвичай R є системою дійсних чисел або її підсистемою, хоча це можуть бути і будь-які символи нечислової природи. Назвемо *допустимими перетвореннями шкали* такі перетворення, відносно яких шкала є інваріантною, тобто при застосуванні цих перетворень не змінює свій тип. Нехай G – група допустимих перетворень шкали. Назвемо *шкалою* упорядковану четвірку $<A; R; f; G>$. Відповідно до сучасних уявлень, внутрішньою характеристикою шкали виступає саме група G, а f є лише прив'язкою шкали до конкретної ситуації вимірювання.

Формування теоретичного конструкту і процес вимірювання. Розглянемо на прикладі процес формування теоретичного психологічного конструкту і як це веде до вимірювання. Припустимо, що шкільний психолог спостерігає за поведінкою дітей під час їхніх ігор. Він помічає, що дехто з дітей постійно намагається управляти діями інших. Після спостережень, виконаних протягом тривалого часу і в різних ситуаціях, наш психолог може дати спеціальну назву такій поведінці, наприклад, «соціальне домінування». Цим самим він винайшов *теоретичний конструкт*, який проявляється у різних схожих за змістом способах поведінки дітей. Поняття конструкт вживається нами для позначення ряду схожих між собою типів поведінки (проявів психічної чи інтелектуальної сфери людини). Конструкти є «цеглинами» у загальній теорії поведінки людини. Теорія потрібна для того, щоб уміти передбачати і при необхідності впливати на поведінку особи.

Але винайдення конструкту ще не є його вимірюванням. Перед тим, як здійснювати вимірювання, потрібно визначити правила відповідності між множиною різних типів поведінки дітей і теоретичним конструктом. Цей процес називають *операційним визначенням*. У нашому прикладі, психолог повинен визначити, які типи

поведінки дітей в спостережуваних умовах є проявами домінування. Далі психолог повинен створити план отримання зразків таких типів поведінки в стандартних ситуаціях і фіксування цих зразків у деякому стандартизованому форматі для кожної дитини окремо. Цим самим психолог виготовляє інструмент для вимірювання соціального домінування, який ми будемо називати тестом. Взагалі кажучи, *тест* можна розуміти як стандартну процедуру отримання зразків поведінки в межах визначеної області. Якщо цією визначеною областю є навчальні досягнення, то тест має перевіряти *оптимальні дії* опитуваних, і їх спеціально налаштовують на якнайкраще проявлення досягнень під час тестування. Якщо досліджуваною областю є ставлення досліджуваних до чогось, чи їх інтереси, реакції на певні ситуації, то тест має перевіряти *типову поведінку*.

Процес отримання зразків поведінки досліджуваних може відбуватися у різних формах. В одному випадку це можуть бути записи спостережень, в іншому – використання приготованих заздалегідь переліків поведінки (тестових завдань). Вимірювання відбувається тоді, коли результати тестування за допомогою визначених правил переводяться у числа. Наприклад, вимірювання виникає під час підрахунку актів соціального домінування, які проявила дитина протягом 10-хвилинного спостереження за її поведінкою в спеціально створених умовах, або при підрахунку правильних відповідей на завдання тесту, який виконали учні класу на уроці хімії. Результатом вимірювання є висновок фахівця щодо рівня теоретичного конструкту, який проявляє кожен опитуваний. У випадку, коли таким фахівцем є педагог, передбачається, що після визначення результату вимірювання цей результат буде використаний ним (або самим опитуваним, чи батьками опитуваного) для прийняття *педагогічного рішення*.

Проблеми вимірювання психологічних конструктів. Оскільки психологічні конструкти – це абстракції, які не можуть оцінюватися безпосередньо (є *латентними*), створення інструментів вимірювання для них завжди пов’язане з рядом проблем. Виділимо п’ять основних проблем.

1. *Не існує одного універсального підходу до вимірювання певного конструкту*. Оскільки психологічний конструкт спостерігається лише через певні типи поведінки, то два теоретики можуть

запропонувати суттєво різні його операційні визначення. Нехай, наприклад, потрібно визначити здатність учня ділити числа у стовпчик. З цією метою можна дати учню ряд завдань на ділення у стовпчик; цієї ж мети можна досягти, якщо попросити учня пояснити, як потрібно виконувати ділення у стовпчик; можна також запропонувати йому знайти помилки у вирішених задачах. З різних операційних визначень випливають різні вимірювальні процедури.

2. *Психологічні вимірювання завжди базуються на обмежених вибірках спостережень*. Для нашого прикладу з діленням, ми не можемо дати учню розв'язати всі можливі задачі з цієї теми. Необхідно визначитися, скількох задач буде достатньо, щоб їх успішне розв'язання адекватно демонструвало вміння ділити у стовпчик і вся предметна область була охоплена.

3. *Вимірювання психологічних конструктів завжди супроводжуються помилками*. Навіть якщо дати одному учню один і той же набір задач, але в різний час, результати навряд чи будуть точно збігатися, оскільки дії учня зажди супроводжується такими важко передбачуваними ефектами як втома, неуважність, забудькуватість, спробами вгадування тощо. Ще більших розбіжностей у результатах слід очікувати, якщо один і той же конструкт вимірюється різними інструментами. Таким чином, завжди актуальною є питання про оцінку похибки вимірювання.

4. *Не існує природного нуля та одиниці вимірювання на обраній шкалі*. Чи означає нуль розв'язаних учнем під час тестування задач на ділення у стовпчик, що у нього відсутні знання з цієї теми? Нехай Володимир розв'язав 5 задач, Марія – 10, а Ольга – 15. Чи означає це, що відмінність у здатності ділити у стовпчик у Володимира і Марії така сама, як у Марії і Ольги? Практично завжди можна стверджувати, що первинні бали (тобто простий підрахунок розв'язаних завдань) не володіють всіма необхідними ознаками міри, і ці бали потрібно трансформувати за допомогою спеціальних процедур у обрану шкалу вимірювання.

5. *Психологічний конструкт не може визначатися лише в термінах операційного визначення, він також повинен проявлятися у зв'язках з іншими конструктами.* Цей другий рівень визначення конструкту дає змогу інтерпретувати результати вимірювання, і ця обставина має надзвичайно велике практичне значення.

2. БАЗОВІ ПОНЯТТЯ СТАТИСТИКИ В ТЕСТУВАННІ

Тестування як вид вимірювання спирається на теорію математичної статистики. У цьому курсі передбачається, що студенти вже вивчали нормативний курс теорії ймовірностей та математичної статистики. Мета цього параграфу – нагадати читачу базові поняття математичної статистики та прив'язати ці поняття до потреб теорії тестування.

Далі розглядатимемо наступну типову ситуацію. Нехай група учнів чи студентів (далі всіх називатимемо учнями) пройшла тест з певної дисципліни, наприклад, біології. Нехай за правильну відповідь на кожне завдання тесту учень отримує певну визначену наперед кількість балів, за неправильну відповідь – 0 балів. Загальним попереднім результатом тестування учня є сума балів, отриманих ним за всі його відповіді. Очевидно, що у різних учнів сума балів може відрізнятися, і вона є наперед невідомою. З точки зору викладача ця сума є випадковою величиною.

З іншого боку, суть процедури тестування полягає в тому, щоб ця сума балів якимось чином відображала вираженість в учня риси або конструкту, що вимірюється (наприклад, рівень успішності засвоєння шкільного курсу біології).

Для цього викладач повинен знати, наскільки тест є валідним (тобто адекватним за рядом важливих аспектів), наскільки він є надійним, яка величина похибки вимірювання тощо. Для відповіді на подібні запитання якраз і використовується апарат математичної статистики.

Вибірковий метод. Найбільш повно поняття і факти математичної статистики використовуються при широкомасштабному стандартизованому тестуванні. У цьому випадку вважається, що тест розробляється для надійного і валідного тестування великої за чисельністю категорії учнів, наприклад, випускників загальноосвітніх середніх шкіл України. Така велика сукупність учнів називається *генеральною сукупністю*, або, простіше, *популяцією*. Під час розробки тесту неможливо визначити і перевірити його характери-

стики на всіх об'єктах популяції. Ця робота проводиться лише на відносно невеликій групі її представників, яка називається *вибіркою*. Отримані на вибірці характеристики тесту можуть, з певною долею достовірності, вважатися справедливими для всієї популяції, якщо вибірка є *репрезентативною*, тобто вона правильно, без спотворень, представляє популяцію щодо вимірюваного конструкту. Зауважимо, що завдання конкретного тесту є, в свою чергу, вибіркою з генеральної сукупності всіх можливих тестових завдань, придатних для вимірювання даної риси чи конструкту.

Загальноприйнятим методом отримання репрезентативної вибірки є *випадковий відбір* представників популяції. Але популяція може бути неоднорідною, тобто складатися з менших популяцій, які істотно відрізняються між собою щодо обставин, які впливають на вираженість риси чи конструкту. Ці менші популяції називають *стратами*. Наприклад, популяція всіх випускників середніх шкіл України може поділятися на страти за типом населеного пункту, у якому розташована школа (місто, село, селище), або за типом школи (звичайна, ліцей, гімназія), або за роком випуску. В подібних випадках для отримання репрезентативної вибірки слід подбати, щоб у вибірці були представлені всі страти у тих долях, у яких вони представлені в генеральній сукупності.

За рівних інших умов, вибірка дозволяє тим точніше визначити характеристики тесту, чим більший об'єм (кількість об'єктів) вона має. Це можна пояснити на такому прикладі: якщо в пологовому будинку одного дня народилося вдвічі більше дівчат, ніж хлопчиків, то це не можна вважати характерним для популяції всіх новонароджених, адже ми знаємо, що у популяції хлопчиків і дівчаток народжується приблизно порівну. Тим не менше, подібна «аномалія» досить часто трапляється у невеликих пологових будинках, де зазвичай народжується лише кілька дітей за день. У великих пологових будинках, де дітей народжується за день десятки або сотні, випадок співвідношення дівчаток до хлопчиків, рівне 2:1, є практично неможливим, натомість, воно зазвичай є ближчим до «правильного» співвідношення 1:1.

З іншого боку, з двох вибірок однакового об'єму, у яких вимірюються різні конструкти, більш інформативною є та, чий конструкт в популяції є менш мінливим. Наприклад, якщо тест перевіряє рівень успішності з окремої теми деякої дисципліни, то для

отримання висновків щодо якості тесту потрібна загалом менша вибірка, ніж для тесту з усієї дисципліни. Пізніше навчимося визначати необхідний для заданих характеристик якості вимірювання об'єм вибірки.

Таблиці частот та діаграми частот. Надалі користуватимемося наступним прикладом. Нехай 50 учнів склали тест з певної навчальної дисципліни. Тест складається з 10 завдань. Відповіді учнів на кожне завдання оцінювалася за *дихотомічною шкалою* (1 бал за правильну відповідь, 0 балів за неправильну). Всі результати зазвичай оформляються у вигляді матриці (таблиці) результатів, у якій кожен окремий рядок містить результати відповідей одного учня на всі завдання тесту, а кожен окремий стовпець – результати відповідей всіх учнів на одне завдання тесту (таблиця 1.1).

Таблиця також містить у крайньому правому стовпці та нижньому рядку суми балів. Ці суми слід розглядати як змінні – реалізації відповідних випадкових величин. Числа у крайньому правому стовпці – це так звані «сирі» бали учнів, отримані ними при проходженні всього тесту. Позначимо цей стовпець-змінну літерою X. Оскільки кожне значення цієї змінної є сумою набраних відповідним учнем балів, тобто сумою нулів і одиниць, то вона може набувати значень від 0 (всі відповіді учня неправильні) до 10 (всі відповіді правильні), всього 11 різних значень. Зокрема, учень за номером 1 відповів правильно лише на половину завдань і отримав відповідно 5 балів, а учень №33 відповів правильно на всі завдання тесту і отримав максимальні 10 балів.

Таблиця 1.1. Матриця результатів тестування

Номер учня	**Бали за завдання 1-10**										**Сума (X)**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
1	0	0	0	1	1	1	0	0	1	1	**5**
2	0	0	0	0	0	0	0	0	0	1	**1**
3	0	0	0	0	1	1	1	1	1	1	**6**
4	0	0	0	1	1	1	0	1	1	1	**6**

5	0	1	0	1	1	0	1	1	1	1	**7**
6	1	0	0	1	1	1	0	1	1	0	**6**
7	0	0	0	0	0	0	0	0	0	1	**1**
8	0	1	0	1	1	1	1	1	1	1	**8**
9	0	0	0	0	1	1	0	1	1	1	**5**
10	0	0	0	1	1	1	1	0	0	1	**5**
11	0	0	0	1	1	1	1	0	0	1	**5**
12	0	0	0	0	1	1	0	0	1	1	**4**
13	0	1	0	1	0	0	1	1	0	1	**5**
14	0	0	0	1	1	1	1	1	0	1	**6**
15	0	0	0	1	1	1	1	0	1	1	**6**
16	0	0	0	0	1	1	1	1	1	1	**6**
17	0	1	0	1	1	1	1	1	1	1	**8**
18	0	0	0	1	1	1	1	0	1	0	**5**
19	0	0	0	1	1	0	1	1	1	1	**6**
20	0	0	0	1	1	1	1	1	1	1	**7**
21	0	0	0	1	0	0	0	1	0	1	**3**
22	0	0	0	1	1	1	1	1	1	1	**7**
23	0	0	0	1	0	0	0	1	1	1	**4**
24	0	0	1	0	1	1	1	1	1	1	**7**
25	0	0	0	0	1	0	1	1	1	1	**5**
26	0	0	1	1	0	1	1	1	1	1	**7**
27	0	0	0	1	1	1	1	1	1	1	**7**
28	0	0	0	1	0	1	1	1	1	1	**6**
29	0	0	0	0	1	1	1	1	1	1	**6**
30	0	1	1	1	1	1	1	1	1	1	**9**
31	0	0	0	0	0	0	1	0	1	0	**2**
32	0	0	0	0	1	1	1	1	1	1	**6**
33	1	1	1	1	1	1	1	1	1	1	**10**
34	0	0	1	1	1	1	1	1	1	1	**8**
35	0	1	1	1	1	0	1	1	1	1	**8**

36	0	1	1	1	1	1	1	1	1	1	**9**
37	0	0	1	0	1	1	1	1	1	1	**7**
38	0	0	0	1	0	1	1	1	1	1	**6**
39	0	0	0	1	1	1	1	1	1	1	**7**
40	0	0	1	1	1	1	1	1	1	1	**8**
41	0	0	0	0	0	0	0	1	1	1	**3**
42	0	0	0	0	1	1	1	1	1	1	**6**
43	0	0	1	1	1	1	1	1	1	1	**8**
44	0	0	1	1	1	1	1	1	1	1	**8**
45	0	1	1	1	1	1	1	1	1	1	**9**
46	0	1	1	1	1	1	1	1	1	1	**9**
47	0	0	1	0	1	1	1	1	1	1	**7**
48	0	0	0	1	0	1	1	1	1	1	**6**
49	0	0	0	1	1	1	1	1	1	1	**7**
50	0	0	0	0	0	0	1	1	1	1	**4**
Сума (*Y*)	**2**	**10**	**14**	**34**	**38**	**38**	**40**	**41**	**43**	**47**	**307**

Наскільки сумарний результат кожного з учнів є типовим для даної вибірки учнів? Іншими словами, як часто зустрічається кожне з можливих значень змінної *X* у стовпці результатів? Відповідь на це питання міститься у другому стовпці *таблиці частот* (таблиця 1.2).

Таблиця 1.2. Таблиця частот

X	**Частоти *f(X)***	**Накопичені частоти *cf(X)***	**Відносні частоти *p(X)***	**Накопичені відносні частоти *cp(X)***
0	0	0	0	0
1	2	2	0,04	0,04
2	1	3	0,02	0,06
3	2	5	0,04	0,1
4	3	8	0,06	0,16
5	7	15	0,14	0,3

6	13	28	0,26	0,56
7	10	38	0,2	0,76
8	7	45	0,14	0,9
9	4	49	0,08	0,98
10	1	50	0,02	1

З таблиці видно, що, наприклад, результат 0 балів не зустрічається у вибірці жодного разу, а результат у 6 балів зустрічається найбільш часто (13 разів), тобто є якоюсь мірою типовим для даної вибірки. Ми не даремно тут кажемо «якоюсь мірою». Припустимо, що в деякій іншій вибірці учнів результат 6 балів зустрічається 14 разів, а всі інші 36 результатів розподілені так: 0 балів, 1 бал, 2 бали – по 12 разів, решта – по 0 разів. У такому випадку справедливо буде вважати тест загалом важким для учнів, а результат найбільш частий результат 6 балів – не дуже характерним. Зауважимо, що подібний розподіл результатів тестування був би, в загальному випадку, дуже підозрілим. Він міг би вказувати, наприклад, на сильну неоднорідність вибірки. Скажімо, так могло б бути, якби тест перевіряв знання математики за 5 клас, а вибірка складалася з 36 першокласників і 14 п'ятикласників. Для нашого ж прикладу результат тестування у 6 балів є більш типовим, оскільки близькі результати, зокрема 5 і 7 балів – теж зустрічаються у вибірці з подібними частотами.

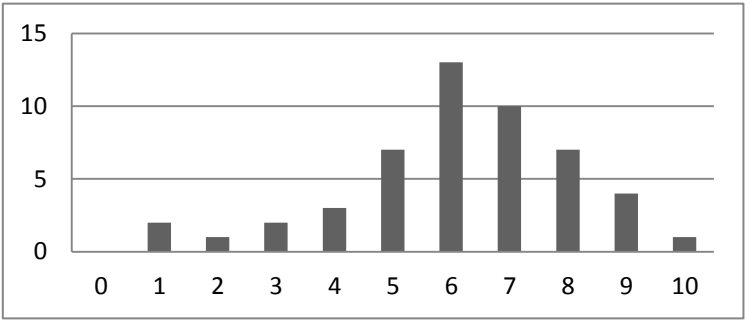


Рис. 1.1. Гістограма частот

Для кращого розуміння результатів тестування частоти $f(X)$ з таблиці 1.2 корисно зобразити у вигляді гістограми (стовпчикової діаграми). Для кожного значення суми балів X, відкладеної вздовж горизонтальної осі, будуємо стовпець, висота якого відповідає частоті цього результату (рис. 1.1).

З малюнка, зокрема, видно, що тест для даної вибірки учнів був загалом нескладним, оскільки найбільші частоти зосереджені в області значень змінної X від 5 до 9. Загалом 41 учнів з 50 отримали за тест від 5 до 9 балів.

Четвертий стовпець таблиці 1.2 містить *відносні частоти* $p(X)$. Для отримання відносної частоти потрібно кожну з частот поділити на об'єм вибірки (у нашому прикладі – на 50). Відносна частота пов'язана з теоретичним поняттям *ймовірності*: якщо вибірка учнів є репрезентативною, то для кожного нового учня з цієї ж популяції ймовірність отримати певну кількість балів за тест приблизно дорівнює відносній частоті цієї кількості балів у вибірці. Наприклад, ймовірність того, що новий учень отримає за тест 7 балів, дорівнює приблизно 0,2.

Числові вибіркові характеристики. Для описання результатів тестування використовуються числові характеристики, які ще називають описовими статистиками. Кожна окрема характеристика – це число, яке характеризує усю вибірку у відповідному аспекті. Всі числові характеристики можна поділити на три групи: міри положення, міри мінливості, міри форми. Далі розглянемо лише деякі, найважливіші, характеристики.

Міри положення, або, як їх ще називають, міри центральної тенденції, вказують на найбільш характерні значення випадкової величини. До цієї групи характеристик відносяться мода, медіана, середнє, а також більш загальне поняття процентиля.

Мода (англ. *mode*) – це значення випадкової величини, яке зустрічається у вибірці найбільш часто. Для нашого прикладу модою є значення $X = 6$, оскільки цей результат зустрічається найбільш часто – 13 разів. Вище ми вже зазначали, що модальне значення в одних випадках добре характеризує типові значення змінної, в інших випадках – гірше. З причин, які розглядатимемо далі, нормо-орієнтований тест слід вважати добре збалансованим по трудності завдань, якщо модальним є значення змінної, близьке до середнього. Так, для нашого прикладу це повинне було б бути

значення, близьке до $X = 5$. Іншими словами, слід було б очікувати, що середній результат у 5 балів набере найбільша кількість учнів.

Медіана (англ. *median*) – це таке значення випадкової величини X, яке ділить вибірку на дві частини приблизно порівну за кількістю об'єктів так, щоб у одній частині опинилися об'єкти із меншими за медіану або рівними їй значеннями, а в іншій частині – з більшими значеннями. Наприклад, якщо у класі 21 учень, то медіана їх зросту дорівнює зросту 11-го учня у вишикуваній за зростом шерензі учнів. Для нашого прикладу тестування 50 учнів медіанним слід вважати результат $X = 6$. Знайти медіану допомагає третій стовпець таблиці 1.2 – стовпець *накопичених частот* $cf(X)$, або п'ятий стовпець цієї таблиці – стовпець *накопичених відносних частот* $cp(X)$. Числа у цих стовпцях отримуються послідовним додаванням кожного нового значення з тих стовпців, які знаходяться зліва (відповідно, частот і відносних частот), до суми попередніх значень. Так, для значення $X = 3$ накопчена частота дорівнює $0 + 2 + 1 + 2 = 5$. Слід пам'ятати, що для відшукання накопичених частот значення змінної X повинні бути розташовані у таблиці за зростанням. Медіанне значення змінної знаходиться у тому рядку таблиці, для якої значення накопиченої частоти досягло 25 (половина вибірки), або значення відносної накопиченої частоти досягло 0,5.

Медіана є частинним випадком *процентиля*: p-й процентиль – це таке значення змінної, яке ділить вибірку так, що p відсотків об'єктів вибірки мають значення, менше або рівне даному. У нашому прикладі учень, який отримав за тест 9 балів, відповідає 98-му процентилю (значення 0,98 у стовпці накопичених відносних частот), що означає, що 98 відсотків учнів склали тест з результатом, не вищим від його результату. При нормоорієнтованому тестуванні результати інколи можуть повідомлятися учням саме у вигляді процентилів. Особливо інформативним є такий підхід у випадку, якщо до нього також додаються інші дані, які характеризують вибірку в цілому, наприклад, у якій школі (спеціалізованій чи звичайній, сільській чи міській) навчаються учні.

Медіана є 50-м процентилем. Вживаються також такі поняття, як *квартилі* – 25-й та 75-й процентилі.

Медіана використовується частіше для описання змінних, виміряних у порядковій шкалі. Для метричних змінних більш інформативним може виявитися поняття середнього.

Середнє вибіркове (англ. *mean*) – це середнє арифметичне усіх значень змінної:

$$\overline{\mathrm{X}} = \frac{\sum_{i=1}^{N} X_i}{N},$$

де N – об'єм вибірки. У нашому прикладі $N = 50$,

$$\bar{X} = \frac{0 \times 0 + 1 \times 2 + 2 \times 1 + \cdots + 9 \times 4 + 10 \times 1}{50} = \frac{307}{50} = 6{,}14.$$

Середнє вибіркове є *оцінкою* теоретичного значення математичного очікування випадкової величини для всієї популяції. Чим більш репрезентативною є вибірка, тим точнішою є ця оцінка.

Слід обережно трактувати значення середнього вибіркового. В літературі часто згадується такий жартівливий приклад: середня температура пацієнтів лікарні може дорівнювати 36,7, проте це зовсім не означає, що пацієнти здорові, і цю середню температуру не можна трактувати як показник успішності роботи лікарів. Середнє вибіркове зазвичай розглядається у парі з характеристикою мінливості змінної – статистичною дисперсією або середнім квадратичним відхиленням.

Статистична дисперсія (англ. *variance)*є характеристикою, яка входить до групи мір мінливості змінної. Ця характеристика є середнім арифметичним квадратів відхилень значень змінної від середнього вибіркового:

$$D(X) = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}.$$

Тобто значення дисперсії показує, як часто і на скільки сильно відхиляються значення змінної від свого центра. Відхилення тут підносяться до квадрату для того, щоб усі доданки були невід'ємними, тобто щоб відхилення з різними знаками не компенсували одне одного. Більш природним було б використання модулів відхилень замість їх квадратів, однак використання модулів є

більш складним. Чим більшим є розсіювання значень змінної навколо середнього, тим більшою є дисперсія. Якщо змінна набуває лише одного значення, то воно ж і є центром, відхилення не спостерігаються, і дисперсія у цьому випадку дорівнює нулю. Якщо змінна набуває більше ніж одного значення, дисперсія завжди є додатною. Розглянемо для прикладу три учнівські класи, які отримали підсумкові оцінки з української мови за 12-бальною шкалою. Нехай у класі *A* всі учні отримали оцінку 8, у ласі *B* – половина учнів отримали 7, інша половина – 9, у класі *C* – половина отримали 6, інша половина – 10. Середня оцінка у кожному з цих класів однакова – 8 балів. Проте очевидно, що за рівнем успішності з предмету окремих учнів ці класи істотно відрізняються. В оцінках класу *A* ніякої мінливості не спостерігається, дисперсія цих оцінок дорівнює нулю. У класі *B* дисперсія вже не нульова, а у класі *C* вона є більшою, ніж у класі *B*.

Статистична дисперсія є оцінкою теоретичної дисперсії змінної для всієї популяції. Проте ця оцінка є *зміщеною*: її математичне очікування не дорівнює теоретичній дисперсії. Виправлену (незміщену) оцінку отримаємо, якщо у знаменнику формули для статистичної дисперсії замінимо N на $N-1$:

$$s^2 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}.$$

У нашому прикладі тестування 50 учнів отримаємо таке значення незміщеної вибіркової дисперсії:

$$s^2 = \frac{0 \times (0-6{,}14)^2 + 2 \times (1-6{,}14)^2 + \cdots + 1 \times (10-6{,}14)^2}{50-1} \approx 3{,}9188.$$

Оскільки у формулі для обчислення дисперсії використовується операція піднесення до квадрату, отримуємо величину, обчислену в квадратних одиницях. Для того, щоб повернутися до тієї ж самої розмірності, яку має вимірювана змінна, візьмемо корінь квадратний з дисперсії:

$$s = \sqrt{s^2}.$$

Ця величина називається *середнім квадратичним*, або *стандартним, відхиленням* (англ. *standard deviation*) випадкової величини *X*. Для нашого прикладу маємо

$$s \approx \sqrt{3{,}9188} \approx 1{,}98.$$

Оскільки операція добування кореня квадратного з невід'ємного числа є монотонним перетворенням, стандартне відхилення має таку ж інтерпретацію, як і дисперсія: воно є мірою мінливості змінної у вибірці.

До групи *характеристик форми* розподілу змінної у вибірці відносяться, в першу чергу, асиметрія і ексцес. Обидві ці характеристики використовуються для описання розподілів, близьких до нормального, тому розглянемо їх пізніше.

Нормальний розподіл. Нормальний, або Гаусів, закон розподілу випадкової величини відіграє особливу роль в теорії і практиці вимірювань. Це пов'язано з тим, що ті випадкові величини, які є сумою багатьох випадкових величин (іншими словами, сформовані під впливом багатьох випадкових факторів), за умови, що вплив кожного з доданків на всю суму не є визначальним, має розподіл, близький до нормального. Оскільки ті випадкові величини, які зустрічаються в природі, на виробництві чи в інших сферах людської діяльності, якраз і формуються під впливом багатьох факторів, їх розподіли часто виявляються близькими до нормального. Зокрема, слушно вважати, що рівень інтелекту розподілений для популяції людей за нормальним законом. Так само з великою мірою впевненості можна стверджувати, що рівень навчальних досягнень учнів з деякого предмету теж добре описується нормальним законом розподілу. Разом з тим, слід пам'ятати, що нормальний розподіл – це лише математична модель, яка в одних випадках добре узгоджується з спостережуваними на практиці величинами, в інших випадках – погано.

Нормальним законом розподілу ймовірностей випадкової величини називається розподіл, щільність якого виражається формулою

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

де μ і σ – параметри розподілу (відповідно, математичне очікування і стандартне відхилення). На малюнку 1.2 схематично зображено графік щільності нормального розподілу.

Щільність розподілу – це функція, графік якої разом з віссю абсцис обмежує площу, рівну ймовірності того, що реалізація випадкової величини потрапить в результаті випробування у заданий інтервал.

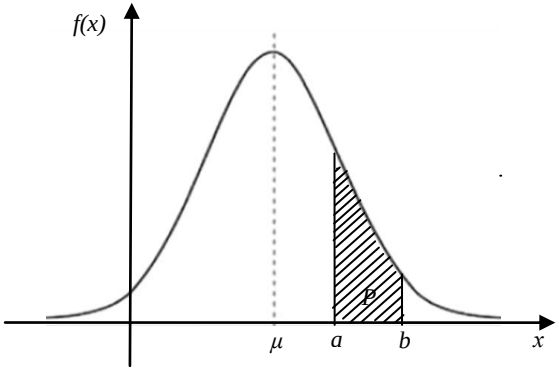


Рис. 1.2. Щільність нормального розподілу

Площа заштрихованої області на рисунку 1.2 дорівнює ймовірності того, що реалізація випадкової величини з таким розподілом потрапить у інтервал $[a, b]$. Очевидно, що ймовірність потрапляння нормально розподіленої випадкової величини в інтервал заданої довжини є найбільшою, якщо цей інтервал є симетричним відносно центра розподілу μ. Тобто реалізації нормально розподіленої випадкової величини зосереджені найбільше поблизу центра. Чим більше відрізняється значення випадкової величини від μ, тим рідше воно трапляється. Іншими словами, люди з середнім зростом, чи з середнім рівнем інтелекту зустрічаються частіше, а з відхиленнями від середнього в ту чи іншу сторону – тим рідше, чим більшим є це відхилення. Зокрема, слід очікувати, що учнів з середнім рівнем навчальних досягнень з певного предмету має бути

більше, ніж учнів з слабким або, навпаки, високим рівнем досягнень.

Варто пам'ятати такі значення ймовірностей потрапляння нормально розподіленої випадкової величини X у заданий симетричний відносно центра розподілу інтервал:

$$P\{X \in [\mu - \sigma, \mu + \sigma]\} \approx 0{,}68.$$

$$P\{X \in [\mu - 2\sigma, \mu + 2\sigma]\} \approx 0{,}95.$$

$$P\{X \in [\mu - 3\sigma, \mu + 3\sigma]\} \approx 0{,}997.$$

Тобто приблизно 68% реалізацій X відрізняються від середнього μ не більше, ніж на σ, 95% - не більше ніж на 2σ. І майже напевне (у 99,7% випадках) відхилення буде не більшим, ніж на 3σ. Останнє твердження називають «правилом трьох сигм».

Важливою властивістю нормального розподілу є той факт, що його математичне очікування, мода та медіана збігаються.

Як ми вже зазначали, нормальний розподіл є у багатьох практичних застосуваннях моделлю, яка більшою чи меншою мірою описує реальні дані. Відхилення реальних даних від нормального закону описується, зокрема, такими мірами форми розподілу, як асиметрія і ексцес.

Асиметрія (точніше, коефіцієнт асиметрії, англ. *skewness*) обчислюється за формулою:

$$As = \frac{E(X - EX)^3}{\sigma^3},$$

де символ E означає математичне очікування. Величина у чисельнику – це центральний момент третього порядку (математичне очікування куба відхилення. Дисперсія є центральним моментом другого порядку). Виправленою оцінкою асиметрії за даними вибірки є величина

$$As = \frac{N}{(N-1)(N-2)} \frac{(\sum_{i=1}^{N}(X_i - \bar{X})^3)}{s^3}.$$

Значення асиметрії дорівнює нулю, якщо щільність розподілу симетрична відносно центра (математичного очікування). Зокрема, якщо випадкова величина розподілена нормально, то асиметрія дорівнює нулю. Якщо асиметрія додатна, то це означає, що правий хвіст розподілу є довшим за лівий (рис. 1.3), а якщо асиметрія від'ємна – то, навпаки, лівий хвіст довший.

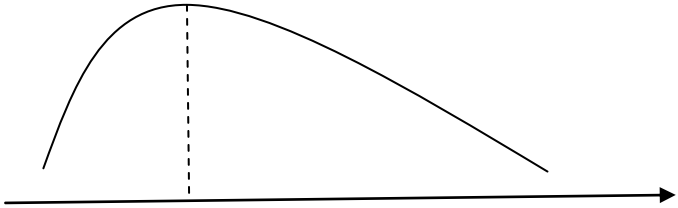


Рис. 1.3. Випадок додатної асиметрії розподілу

Якщо асиметрія результатів тестування популяції учнів з нормальним розподілом рівня успішності додатна, це може означати, що тест є надто складним для учнів, якщо асиметрія від'ємна – то, навпаки, надто легким.

У нашому прикладі тестування 50 учнів гістограма розподілу частот (рис. 1.1) показує, що крива, яка огинає гістограму, має вершину, зсунуту вправо від середнього значення частоти, і лівий хвіст розподілу є довшим. Отже, слід очікувати, що асиметрія розподілу результатів тестування є від'ємною. Дійсно, підставивши значення у формулу для вибіркової оцінки асиметрії, знайдемо, що вона дорівнює приблизно $-0{,}663$.

Ексцес (англ. *kurtosis*) є мірою гостровершинності розподілу у порівнянні з нормальним. Якщо розподіл більш крутий, ніж нормальний, ексцес є додатним. Якщо розподіл більш пологий, то ексцес від'ємний. Величина теоретичного коефіцієнта ексцесу обчислюється за формулою:

$$Ex = \frac{E(X - EX)^4}{\sigma^4} - 3.$$

Перший доданок є відношенням центрального моменту четвертого порядку до четвертого степеня стандартного відхилення. Для нормального розподілу цей доданок дорівнює 3. Тому у формулу вводиться другий доданок, щоб визначений таким чином коефіцієнт дорівнював для нормального розподілу нулю. Незміщена оцінка ексцесу обчислюється за формулою:

$$Ex = \frac{N(N+1)}{(N-1)(N-2)(N-3)} \sum_{i=1}^{N} \left(\frac{X_i - \bar{X}}{s} \right)^4 - \frac{3(N-1)^2}{(N-2)(N-3)}.$$

Для нашого прикладу тестування 50 осіб (таблиця 1.1) величина вибіркової незміщеної оцінки ексцесу дорівнює приблизно 0,619, тобто розподіл тестових балів у нашому випадку має більш гостру вершину у порівнянні з нормальним.

Як ми вже зазначали, багато випадкових величин, з яким має справу людина у своїй практичній діяльності, добре моделюються нормальним розподілом. Разом з тим, ми добре розуміємо, що розподіл, наприклад, зросту чоловіків відрізняється від розподілу зросту жінок, який в свою чергу, відрізняється від розподілу рівня інтелекту у жінок. Усі ці розподіли відрізняються значеннями параметрів середнього μ і стандартного відхилення σ.

Як впливають значення параметрів розподілу на положення і форму кривої щільності розподілу? Якщо змінювати лише параметр μ, то форма кривої не змінюватиметься, але сама вона буде зсуватися вздовж горизонтальної осі так, щоб її вершина була у точці з новим значенням μ. Якщо, навпаки, змінювати параметр σ, то крива буде стискатися або розтягуватися вздовж вертикальної осі: при збільшенні значення σ крива буде ставати більш пологою, а при зменшенні – більш крутою. Це стає зрозумілим, якщо пригадати, який зміст має стандартне відхилення: чим більше його значення, тим частіше зустрічаються відхилення від центра, тобто щільність у самому центрі зменшується, зростаючи натомість у більш віддалених точках.

Будь-який нормальний розподіл можна легко перетворити на *стандартний нормальний розподіл*, тобто розподіл з середнім 0 і

стандартним відхиленням 1, за допомогою заміни змінної, яку називають z-перетворенням:

$$z = \frac{x - \mu}{\sigma}.$$

Це перетворення є надзвичайно важливим з двох причин. По-перше, таблиці значень щільності і функції нормального розподілу існують тільки для випадку $\mu = 0$ і $\sigma = 1$. По-друге, для порівняння схожих за природою, але виміряних у різних шкалах, довільно розподілених величин, їх потрібно спочатку привести до єдиної шкали. Як порівняти, наприклад, оцінки з двох предметів, якщо один з них оцінювався за шкалою 100-200 балів, а інший – за шкалою 0-10 балів? Якщо взяти за базову першу шкалу, то оцінки, отримані за другою шкалою, слід перетворити, помноживши їх на 10 (перехід до шкали 0-100) і потім додавши до них 100. У випадку зведення нормальних розподілів вигідно обидва звести до стандартного, оскільки для нього існують таблиці значень щільності та функції розподілу.

Якщо потрібно знайти ймовірність, з якою нормально розподілена випадкова величина потрапить у заданий інтервал $[a,\ b]$, використовується формула:

$$P\{X \in [a, b]\} = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right),$$

де $\Phi(x)$ - функція Лапласа, значення якої можна знайти за спеціальною таблицею. Ця функція відрізняється від функції *стандартного нормального розподілу* (тобто нормального розподілу з середнім 0 і стандартним відхиленням 1) у кожній точці на величину 0,5, але нею користуватися більш зручно, тому що вона є непарною (тобто $\Phi(-x) = -\Phi(x)$) і тому немає потреби заносити у таблицю її значень значення для від'ємних x. Вирази в дужках є z-перетворенням відповідних кінців відрізка. На малюнку 1.4. схематично зображені графіки функції Лапласа та функції стандартного нормального розподілу.

З малюнка видно, що вже при $X = 3$ значення функції Лапласа наближається до 0,5.

Коли потрібно аналізувати вибірку, зокрема, дані тестування групи осіб, то для того, щоб використати нормальний розподіл у якості моделі, необхідно вирішувати, наскільки добре модель підходить, і з якими значеннями параметрів. Гіпотезу про відповідність розподілу вибіркових даних нормальному для заданого рівня значущості слід перевіряти відповідними статистичними методами. Для цього краще скористатися спеціальним програмним забезпеченням.

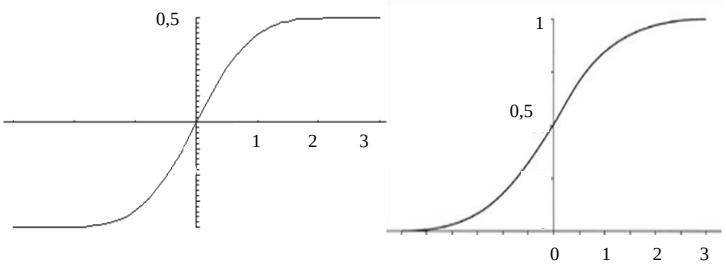


Рис. 1.4. Зліва – функція Лапласа, справа – функція стандартного нормального розподілу

Статистичний зв'язок між змінними. Розглянемо тепер ситуацію, коли в кожному випробуванні спостерігається не одна, а дві випадкові змінні. Наприклад, група учнів складає два тести з різних дисциплін. Тепер поняття вибірки ми не будемо ототожнювати з об'єктами (у нашому випадку – учнями), а з самими змінними – характеристиками об'єктів, які вимірюються. Таким чином, під вибіркою розумітимемо просто сукупність чисел, яка має певний статистичний розподілу, тобто для однієї і тієї ж групи N досліджуваних будемо розглядати дві, за кількістю змінних, вибірки об'єму N.

При одночасному спостереженні двох змінних з'являється принципово нове і надзвичайно важливе питання: чи існує зв'язок між змінними? Припустимо, що перед нами учень, про якого відомо, що він складав два тести – з алгебри та з геометрії. Припустимо також, що нас цікавить результат тестування цього учня з геометрії. Якщо розподіл тестових балів з геометрії для групи, до якої

належить наш учень, відомий, ми можемо на основі цього розподілу робити припущення щодо оцінки цього учня. Наприклад, якщо відомо, що дві третини учнів отримали менше 70 балів, ми з упевненістю в майже 67 відсотків очікуємо, що й наш учень набрав менше 70 балів. Але нехай нам стало відомо, що учень отримав найвищий бал з алгебри. Швидше за все, це змусить нас переглянути свої припущення щодо його оцінки з геометрії: тепер ми з меншою ймовірністю, ніж 67 відсотків, очікуємо, що у нього менше 70 балів, натомість зростає наша віра у те, що його оцінка є вищою. Це відбувається тому, що ми припускаємо наявність зв'язку між оцінками з алгебри та геометрії: чим вищою є оцінка з алгебри, тим вищою вона має бути й з геометрії.

Але чи можемо ми, погоджуючись з існуванням подібного зв'язку, і знаючи оцінку учня з алгебри, зі 100-відсотковою упевненістю назвати його оцінку з геометрії? Очевидно, ні. На обидві оцінки впливає багато випадкових факторів, які ми не можемо врахувати. Хоча ми й очікуємо, що висока оцінка з алгебри означає також високу оцінку з геометрії (і навпаки), цілком ймовірною є, наприклад, ситуація, коли одна з оцінок учня є високою, а інша – низькою, і цьому може бути багато причин. Іншими словами, припускаючи існування зв'язку між змінними, ми, разом з тим, розуміємо, що цей зв'язок не є однозначним, іншими словами, він є ймовірнісним.

Взагалі кажучи, між двома змінними може існувати кілька різних видів зв'язку. Для нас надалі важливо буде чітко розрізняти три види зв'язку: функціональний, статистичний, кореляційний.

Функціональним називається такий зв'язок між змінними X та Y, коли кожному можливому значенню X відповідає одне і тільки одне значення Y. Часто функціональний зв'язок може задаватися математичним рівнянням, наприклад: $Y = 2X + 3$. У цьому прикладі значенню $X = 5$ відповідає єдине значення $Y = 13$. Якби між результатами тестування з двох дисциплін існував функціональний зв'язок, ми, знаючи одну з оцінку, знали б і іншу. Зауважимо, що у такому разі одне з тестувань виявилося б просто зайвим.

В реальних ситуаціях, спостерігаючи дві змінні, кожна з яких є випадковою величиною (а саме такими і є результати двох різних психометричних вимірювань), ми ніколи не побачимо між ними функціонального зв'язку. Втім, це не обов'язково означає,

що зв'язку немає взагалі. Між змінними може існувати *ймовірнісний* (*стохастичний, статистичний*) зв'язок. Тут важливо розуміти, як співвідноситься теорія з практикою, математична модель – з даними спостережень. Розглядаючи рівні навчальних досягнень групи учнів з алгебри і геометрії, цілком природно припустити, що між ними *в теорії* існує функціональний зв'язок у формі, наприклад, лінійної залежності виду $Y = aX + b$, чи квадратичної залежності виду $Y = aX^2 + b$ тощо, але реальні результати вимірювання шляхом тестування показуватимуть відхилення від цього зв'язку в той чи інший бік, унаслідок впливу різноманітних сторонніх випадкових факторів. Також ми в теорії можемо допускати існування зв'язку між рівнями навчальних досягнень з алгебри і української мови, але природно очікувати, що цей зв'язок є *менш тісним*, ніж у випадку алгебри і геометрії, тобто відхилення від функціонального зв'язку між результатами тестування з алгебри і української мови очікуються в середньому більшими. Таким чином, важливою характеристикою статистичного зв'язку є його *сила* (*тіснота*), як міра відхилення від теоретичного функціонального зв'язку.

Розглянемо реальний приклад – результати тестів ЗНО 2010 року з математики (X) і української мови та літератури (Y) п'ятдесяти вступників до фізико-математичного факультету Ніжинського державного університету імені Миколи Гоголя. Для більшої виразності бали з української мови та літератури переведені з шкали 100-200 у шкалу 0-10 балів лінійним перетворенням: від кожної оцінки ЗНО відняли 100 балів і результат поділили на 10. Подамо результати у вигляді таблиці 1.3.

Таблиця 1.3. Результати ЗНО 50 вступників

Укр. мова	**Математика**
10	181
9	174 194 190 187
8	184 180 177 194 174 178 188 173 158 165
7	180 188 174 168 177 171 160 163 168 173 168
6	179 175 169 190 156 170 177 147 158 154
5	162 174 129 166 147 157
4	155 162 158 147 147 125
3	142 140

У таблиці дані згруповані за оцінкою з української мови: один вступник мав 10 з мови і 181 з математики, чотири вступники мали 9 з мови і 174, 194, 190 та 187 з математики відповідно і так далі. Як бачимо для кожного із значень оцінки з мови та літератури існує свій набір оцінок з математики. Цей набір називається умовним розподілом. Наприклад, у передостанньому рядку бачимо умовний розподіл оцінок вступників з математики *за умови*, що оцінка з мови дорівнює 4. Іншими словами, *умовний розподіл* змінної Y за умови $X = a$ – це набір значень, які набула змінна Y при фіксованому значенні змінної $X = a$. Позначимо тепер оцінки кожного вступника точкою на декартовій площині з координатами (українська мова та література, математика) (рис. 1.5).

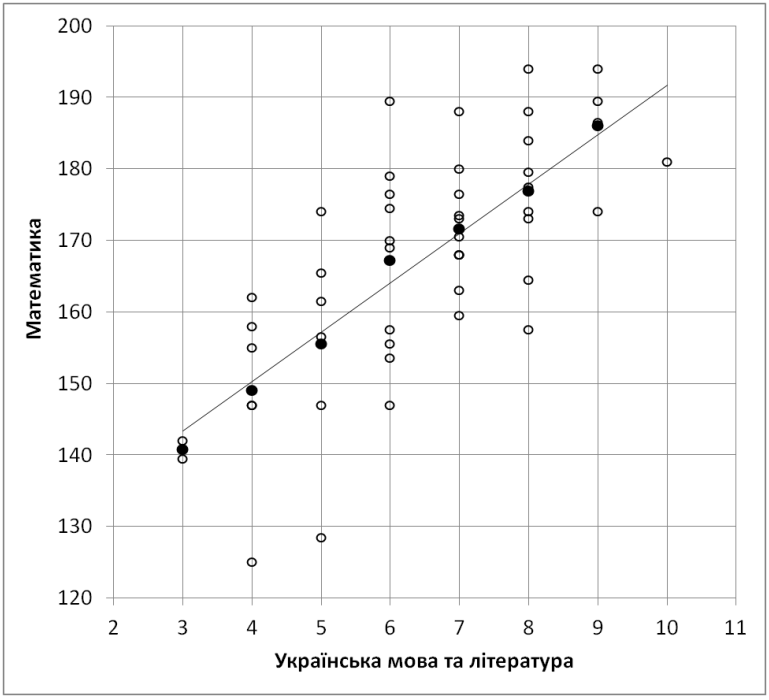


Рис. 1.5. Розподіл оцінок 50 вступників за двома предметами

Для кожної оцінки з української мови та літератури знайдемо середнє значення відповідних оцінок з математики і позначимо його на рисунку чорним кружечком. Як бачимо, ці середні значення розташовані досить близько до деякої прямої. Це дає змогу припустити, що умовні математичні очікування оцінок з математики усієї популяції вступників, до якої належить вибірка, знаходяться на цій прямій, тобто існує теоретична лінійна залежність між оцінками. Сама пряма називається *прямою регресії* змінної Y по змінній X або *лінією тренду*. Теоретичне рівняння прямої регресії Y по X має вигляд:

$$y - EY = r_{xy}\frac{\sigma_Y}{\sigma_X}(x - EX).$$

Тут EX, EY – математичні очікування випадкових величин X та Y; σ_X, σ_Y – їх середні квадратичні (стандартні) відхилення. Нова для нас величина r_{XY} – це коефіцієнт кореляції Пірсона. Ця величина відіграє надзвичайно важливу роль, тому далі розглянемо її детально. Оскільки, володіючи лише вибірковими даними, ми не зможемо дізнатися істинних значень математичних очікувань змінних, їх стандартних відхилень та коефіцієнта кореляції, то замість теоретичного рівняння прямої регресії ми можемо використовувати *вибіркове рівняння*, яке отримаємо, замінивши у наведеній вище формулі теоретичні величини їх вибірковими оцінками:

$$y - \bar{Y} = \rho_{XY}\frac{s_Y}{s_X}(x - \bar{X}).$$

Тут ми замінили математичні очікування їх вибірковими оцінками – середніми арифметичними відповідних вибірок, а стандартні відхилення – вибірковими стандартними відхиленнями. Формула для вибіркової оцінки коефіцієнта кореляції Пірсона:

$$\rho_{XY} = \frac{\sum_{i=1}^{N}(x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{N}(x_i - \bar{X})^2 \sum_{i=1}^{N}(y_i - \bar{Y})^2}.$$

Усі вибіркові оцінки отримують зазвичай за допомогою спеціальних комп'ютерних програм. Наприклад, у середовищі

Microsoft Excel для вибіркового рівняння прямої регресії, записаного у стандартному вигляді $y = ax + b$, для обчислення параметра b, який називається інтерцептом і вказує на точку перетину прямої з віссю Oy, використовується стандартна функція INTERCEPT, а для обчислення параметра a, який є коефіцієнтом нахилу прямої до осі Ox – функція SLOPE. У нашому прикладі отримаємо рівняння прямої:

$$y = 6{,}906x + 122{,}637.$$

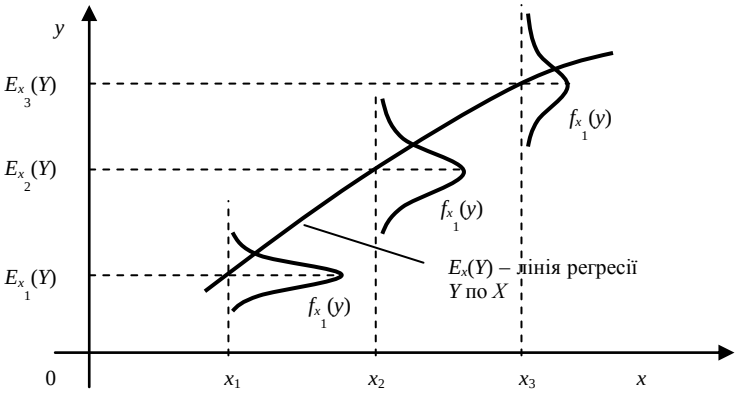
Навіщо потрібна пряма регресії? Знаючи її рівняння, можна прогнозувати оцінку деякого вступника з математики на основі його оцінки з мови та літератури. Припустимо, що вступник має оцінку 8 з мови та літератури. Підставивши це значення у рівняння, отримаємо:

$$y = 6{,}906 \times 8 + 122{,}637 \approx 178.$$

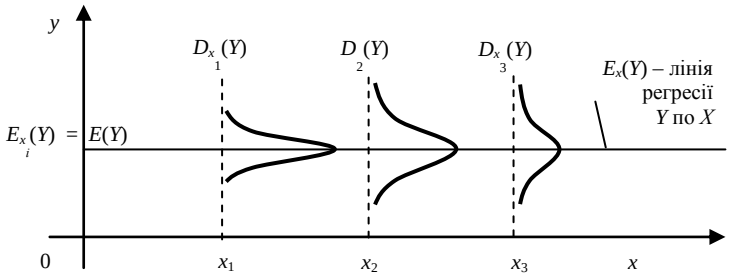
У програмі Excel є спеціальна функція TREND для обчислення прогнозованого значення *Y* за даним *X*.

Таким чином, очікувана оцінка цього вступника з мови та літератури – 178 балів. Зауважимо, що зазвичай реальна оцінка відрізняється від прогнозованої. Точність прогнозу залежить від тісноти статистичного зв'язку між змінними. Розглянемо тепер більш детально величину, яка характеризує силу лінійного зв'язку між змінними – коефіцієнт кореляції Пірсона.

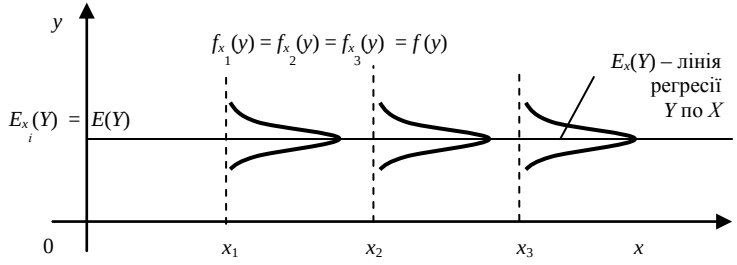
Кореляційний зв'язок між змінними. Коефіцієнти кореляції Пірсона та Спірмена. Перш за все, зауважимо, що статистичний та кореляційний зв'язки – це різні поняття. Відсутність кореляційного зв'язку не означає, що між змінними немає статистичного зв'язку. Розглянемо три різні випадки, зображені на рис. 1.6. У випадку а) при зростанні змінної x в цілому спостерігається також і зростання y, причому із зміною x також змінюються і умовні розподіли y – збільшується дисперсія розподілу. Це свідчить про наявність кореляційного зв'язку між змінними, який, проте, не є лінійним. У випадку б) лінія регресії є горизонтальною прямою, що вказує на відсутність лінійного кореляційного зв'язку (коефіцієнт a у рівнянні прямої дорівнює нулю).



а)



б)



в)

Рис. 1.6. Співвідношення статистичного та кореляційного зв'язку

Проте між змінними все ж спостерігається статистичний зв'язок: із зростанням x змінюються умовні розподіли y. Нарешті, випадок в) вказує на відсутність і кореляційного, і статистичного зв'язку між змінними – умовні розподіли змінної y при зміні x залишаються однаковими.

На рис. 1.7 зображено випадок, коли між змінними існує функціональний зв'язок у вигляді залежності $x = y^2 + 10$, а проте лінія тренду є горизонтальною, тобто лінійна кореляція не спостерігається.

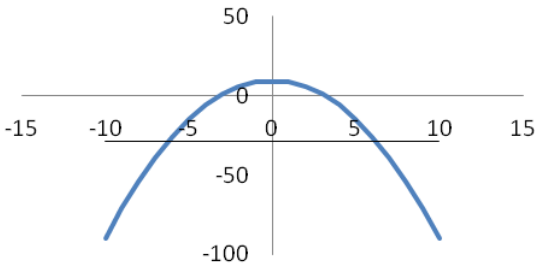


Рис. 1.7. Між змінними існує строга функціональна залежність, але відсутній лінійний кореляційний зв'язок

Кореляційний зв'язок (або, простіше, кореляція) між змінними, найчастіше має лінійну форму. Тому часто, взагалі кажучи, не коректно, говорять, що змінні корелюють між собою, маючи на увазі саме лінійний зв'язок.

Коефіцієнтом кореляції Пірсона називається величина, яка виражається формулою

$$r_{XY} = \frac{E((X - EX)(Y - EY))}{\sigma_X \sigma_Y}.$$

Математичне очікування добутку відхилень X та Y, яке знаходиться у чисельнику – це так звана *коваріація* X та Y. Відхилення $X - \bar{X}$ – це випадкова величина, які отримуються зсувом усіх значень випадкової величини X на число EX вздовж осі. Математичне очікування величини X переміститься при цьому у точку нуль, і

тому відхилення є *центрованою* випадковою величиною, тобто його математичне очікування вже дорівнює нулю.

Величина коефіцієнта кореляції Пірсона вказує як на напрям лінійного зв'язку між змінними, так і на силу (тісноту) цього зв'язку. Це число завжди лежить в інтервалі [-1, 1]. Якщо коефіцієнт додатній, то зв'язок між змінними прямо пропорційний (із зростанням однієї змінної спостерігається в цілому зростання іншої, як у розглянутому вище прикладі). Якщо коефіцієнт від'ємний, то це означає, що між змінними спостерігається обернено пропорційний лінійний зв'язок (із зростанням однієї змінної спостерігається в цілому спадання іншої), у цьому випадку графіком пряма регресії є спадною. Якщо коефіцієнт кореляції за модулем близький до нуля, то зв'язок є слабким. Якщо коефіцієнт кореляції близький до одиниці або мінус одиниці, то зв'язок є сильним. Нарешті, якщо коефіцієнт кореляції дорівнює плюс або мінус одиниці, це означає, що між змінними існує функціональний лінійний зв'язок (усі точки лежать на прямій). Останній випадок на практиці не спостерігається через наявність випадкових впливів на значення змінних. Слід добре розуміти, що вибіркова оцінка коефіцієнта кореляції практично завжди відмінна від нуля. Для малих вибірок значення, наприклад 0,1 може означати, що насправді для усієї популяції об'єктів вимірювання кореляційний зв'язок відсутній, а відмінність отриманого числа від нуля є наслідком обмеженості вибірки. У цьому випадку кажуть, що отримане значення вибіркової оцінки коефіцієнта кореляції не є *значущим*. Значущість отриманої оцінки перевіряється за допомогою спеціальних методів перевірки статистичних гіпотез.

Якщо змінні виміряні у порядковій шкалі, то у якості міри лінійного кореляційного зв'язку між ними використовується *коефіцієнт кореляції Спірмена*. Цей коефіцієнт отримаємо, замінивши у формулі вибіркового коефіцієнта Пірсона значення змінних їх порядковими номерами – рангами. Альтернативною мірою для порядкових змінних є коефіцієнт «тау» Кендалла. Зауважимо, що порядок у обох змінних має бути однаковим – за зростанням вираженості вимірюваної ознаки або за спаданням.

Стандартна похибка вимірювання та довірчий інтервал. Якщо рівняння регресії використовувати для передбачення значення однієї оцінки особи за значенням її іншої оцінки, постає

питання про точність такого прогнозу. Відповісти на нього допомагає поняття стандартної похибки вимірювання. Якщо ми будемо багаторазово робити вибірку одного і того ж об'єму з популяції і для кожної вибірки застосовувати вимірювання, то, очевидно, ми щоразу отримуватимемо різні значення середнього вибіркового, тобто середнє вибіркове є випадковою величиною. Стандартне відхилення цього розподілу називається *стандартною похибкою вимірювання*. Нехай по даному значенню оцінки *x* потрібно спрогнозувати значення оцінки y'. Тоді стандартна похибка прогнозу обчислюється як

$$s_{YX} = s_Y\sqrt{1-\rho_{XY}^2}.$$

Якщо припустити, що похибки прогнозу нормально розподілені навколо кожного значення y' з однаковою умовною дисперсією, то це дає змогу визначити, з якою упевненістю можна стверджувати, що істинне значення оцінки *y* потрапляє у той чи інший окіл оцінки y', який називається *довірчим інтервалом*. Так, з упевненістю (надійністю) приблизно 0,68 (або 68%), істинна оцінка потрапляє в інтервал $y' \pm 1s_{XY}$, а з надійністю приблизно 95% - у інтервал $y' \pm 2s_{XY}$.

Кореляційний та причинний зв'язки. Наявність статистичного чи кореляційного зв'язку нічого не говорить про причинно-наслідкові зв'язки між змінними. У одних випадках зовнішня інформація дозволяє легко вказати, що є причиною, а що наслідком. Наприклад, очевидно, що між часом, який учень щодня витрачає на вивчення математики, та оцінками з математики існує пряма кореляція, причому збільшення часу є причиною зростання оцінки. Змінну-причину називають *незалежною змінною*, а змінну-наслідок – *залежною*.

У інших випадках обидві змінні корелюють внаслідок впливу якоїсь третьої змінної. Сюди можна віднести розглянутий вище приклад кореляції між оцінками вступників з математики та мови і літератури. Нехтування причинно-наслідковими зв'язками може привести до помилок в інтерпретації зв'язку між змінними. У психологічній літературі часто цитується наступний приклад. Деякий

психолог помітив, що між довжиною стопи учнів та їх математичними здібностями існує пряма лінійна залежність. З цього наш психолог зробив сенсаційний висновок: довжина стопи *впливає* на математичні здібності дітей! Насправді ж у дослідженні брали участь діти різного віку. Зрозуміло, що саме збільшення віку є причиною як зростання довжини стопи, так і математичних здібностей дитини. Можуть навіть зустрічатися ситуації, коли між двома змінними існує додатна кореляція, але внаслідок впливу третьої змінної спостережена кореляція є від'ємною!

Усунути ефект впливу третьої змінної Z можна за допомогою так званого *частинного коефіцієнта кореляції*:

$$r_{XY|Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}.$$

Наприклад, якщо $r_{XY} = 0{,}63, r_{XZ} = 0{,}9,\ r_{YZ} = 0{,}7,$ то за цією формулою отримаємо $r_{XY|Z} = 0$, тобто змінні X та Y не корельовані!

Коефіцієнт детермінації. Піднісши коефіцієнт кореляції до квадрату, отримаємо величину, яку називають *коефіцієнтом детермінації*. Ця величина вказує на частку дисперсії залежної змінної, яка обумовлена впливом незалежної змінної. Розглянемо приклад. Різні дослідження вказують на те, що між коефіцієнтом IQ та показниками успішності школярів існує кореляція в межах від 0,5 до 0,7. Отже, коефіцієнт детермінації лежить у межах від 0,25 до 0,49. Тобто можна стверджувати, що дисперсія середнього бала успішності може бути передбаченою за результатами тестування IQ не більше ніж на 25%-49%. Іншими словами, якби ми розглядали учнів лише з певним однаковим показником інтелектуального розвитку, мінливість показника їх успішності була би на 25%-49% меншою у порівнянні з усією популяцією учнів. Зауважимо, що отримані межі коефіцієнта детермінації є досить низькими і це вказує на те, що для прогнозування рівня успішності учнями замало користуватися лише значеннями їх IQ. Для прогнозування краще використовувати множинну регресію, яка розглядає залежність змінної від більше ніж однієї незалежних змінних. Тому, наприклад, для прийому абітурієнтів до ВНЗ в Україні, тобто для про-

гнозування майбутньої успішності навчання студентів, враховуються кілька оцінок зовнішнього незалежного тестування, а також середній бал шкільного атестату. Щоправда, на момент 2012 року кожна з цих оцінок враховується з однаковим ваговим коефіцієнтом, тоді як зрозуміло, що не всі вони однаково добре підходять для прогнозу. Оптимальні вагові коефіцієнти для кількох незалежних змінних можна отримати з рівняння множинної лінійної регресії, для якої розглянуте нами вище рівняння парної регресії є частинним випадком.

Тестова оцінка як сума. Зазвичай на початкових етапах тестування оцінка екзаменованого за тест є простою сумою балів, отриманих ним за кожне з тестових завдань. Також часом тест є *батареєю тестів*, тобто складається з окремих субтестів, кожен з яких призначено для вимірювання деякої окремої якості. В усіх випадках оцінка за тест складається з окремих оцінок і розробнику тесту важливо знати, як залежать статистичні властивості тестової оцінки від властивостей оцінок за окремі завдання чи субтести.

З усіх схем оцінювання окремого тестового завдання важливо виділити *дихотомічну.* Це оцінювання, яке допускає лише дві оцінки – 0 за відповідь, яка підтверджує вимірювану якість, 1 – за відповідь, яка цю якість не підтверджує. Такими як правило є завдання закритого типу з однією правильною відповіддю або завдання з пропущеним словом. Не дихотомічні схеми оцінювання допускають кілька градацій оцінки, наприклад, 0 – за повністю неправильну відповідь, 1 – за частково правильну, 2 – за повністю правильну відповідь. Такі схеми використовуються в завданнях закритої форми з кількома правильними відповідями, зокрема, завданнями на відповідність, а також есе. Для групи екзаменованих можна визначати, як характеризує вимірювану якість кожне з завдань, розглядаючи розподіл оцінок за дане завдання у групі.

Для дихотомічних завдань відношення кількості правильних відповідей до кількості всіх відповідей у групі екзаменованих є одночасно середнім арифметичним і показником *трудності* завдання (позначають як p_j для j-го завдання). Зауважимо, що термін «трудність» є технічним, його зміст прямо протилежний розмовному значенню цього слова – чим більше отримано правильних

відповідей на завдання, тим воно є легшим для екзаменованих, але тим більшою є його трудність як психометрична величина.

Для статистичних характеристик дихотомічних завдань існують формули, які полегшують обчислення, у порівнянні з традиційними.

Дисперсію розподілу відповідей на дихотомічне завдання легко обчислити, помноживши трудність завдання на частку неправильних відповідей (тобто «легкість» завдання $q_j = 1 - p_j$):

$$s_j^2 = p_j q_j.$$

Зокрема, якщо трудність завдання дорівнює 0,5, дисперсія його розподілу дорівнюватиме 0,25, а стандартне відхилення – 0,5.

Для того, щоб обчислити коефіцієнт кореляції між двома дихотомічними завданнями з номерами j і k, зручно скористатися формулою так званого φ-коефіцієнта:

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

де p_{jk} – частка тих екзаменованих, які дали правильні відповіді одночасно на обидва завдання. Хоча цей коефіцієнт і має своє позначення, він є все тим же коефіцієнтом кореляції Пірсона.

Розглянемо тепер особливості оцінки за тест як суми оцінок за завдання.

Нехай тест X складається з двох завдань X_1 та X_2 і його склала група з 10 осіб. Для відшукання середнього значення по кожному завданню потрібно суму балів за це завдання поділити на 10. Для відшукання ж середнього оцінки за весь тест потрібно спочатку додати оцінки за два завдання для кожного екзаменованого, щоб знайти його оцінку за тест, а потім поділити суму усіх оцінок за тест на 10. Таким чином, *середнє значення оцінок за тест дорівнює сумі середніх за кожне завдання*:

$$\bar{X} = \overline{X_1} + \overline{X_2}.$$

За індукцією отримаємо аналогічний результат і для тесту з будь-якою кількістю завдань.

Дисперсія оцінок за тест обчислюється за формулою:

$$s_X^2 = \sum_{i=1}^{N} s_i^2 + 2\sum_{i<j} \rho_{ij} s_i s_j.$$

Згідно з даним раніше означенням, під знаком другої суми стоять коваріації різних пар завдань. Таким чином, кожна додатна кореляція між парою завдань збільшує дисперсію оцінки за тест, а кожна від'ємна – зменшує. Цей факт можна частково перевірити, розглянувши дві вибірки гіпотетичних оцінок десяти учнів за два завдання, з яких одна – це числа 1, 2, …, 10, а друга – ті ж самі числа, але у зворотному порядку: 10, 9, …, 1. Тут є лінійний обернено пропорційний функціональний зв'язок між змінними: $X_2 = 11 - X_1$, тобто коефіцієнт кореляції дорівнює -1. Дисперсії (і стандартні відхилення) вибірок очевидно однакові: $s_1^2 = s_2^2$. Тоді

$$s_X^2 = s_1^2 + s_2^2 - 2s_1 s_2 = 0.$$

Тобто мінливість результатів тестування відсутня. Це й зрозуміло, адже кожен з учнів отримає одну і ту ж оцінку за тест 11 балів. Такий тест був би абсолютно непотрібним, оскільки він не дозволяє диференціювати екзаменованих. З іншого боку, якби оцінки за друге завдання були такими ж, як і за перше, то оцінка за тест коливалася б від 2 до 20, і такий розмах є найбільш можливим для даних множин чисел.

Отримані результати дозволяють зробити ряд важливих висновків. По-перше, збільшення кількості завдань приводить до збільшення мінливості результатів тестування лише у випадку, коли між кожною парою завдань існує додатна кореляція.

По-друге, для забезпечення максимальної мінливості результатів тестування (а отже, й роздільної здатності тесту) між завданнями повинна бути не тільки достатньо висока кореляція, але й трудність завдань має бути близькою до середньої. Наприклад, для завдань з дихотомічним оцінюванням дисперсія $s^2 = pq$ є найбільшою (0,25) при $p = q = 0{,}5$. З іншого боку, слід пам'ятати, що

для диференціації дуже слабких або дуже сильних екзаменованих тест повинен містити, відповідно, деяку кількість дуже простих та складних завдань. Також слід розуміти, що якщо на два завдання кожен з учнів дає однаково правильну або неправильну відповідь, то, взагалі кажучи, одне з цих завдань є лишнім, адже воно нічого не додає до уже наявної інформації про диференціацію екзаменованих, хоча й ідеально збільшує дисперсію оцінок за тест.

По-третє, намагання розробника тесту включити до нього завдання з різних частин цільової області, щоб охопити тестом якомога більше матеріалу, може призвести до слабкої кореляції між відповідями на ці завдання, що може призвести до недостатньо високої мінливості тестових оцінок.

Але збільшення дисперсії результатів тестування хоча й бажане, все ж не є саме по собі показником якості тесту. Перш за все, тест має бути валідним та надійним, і штучне збільшення дисперсії тестових оцінок не повинне погіршувати цих характеристик. Фундаментальні поняття валідності та надійності вимірювання розглянемо пізніше.

3. ПРОЦЕС КОНСТРУЮВАННЯ ТЕСТУ

Природа освітнього, як і будь-якого іншого психометричного вимірювання, зумовлює певну послідовність дій, які необхідно виконати розробникам тесту для того, щоб вимірювання було якісним.

Можна виділити наступні етапи процесу конструювання тесту:

1. Визначення мети вимірювання.
2. Визначення поведінкових характеристик, які визначають вимірювану якість – рису або конструкт, вибірка змісту для тестування.
3. Підготовка специфікацій тесту. Специфікація тесту повинна містити інформацію про пропорції завдань, які представляють визначені на попередньому етапі поведінкові характеристики.
4. Створення початкової множини тестових завдань.
5. Рецензування завдань, їх коригування, якщо необхідно.
6. Проведення попередньої апробації завдань.
7. Апробація завдань на значній репрезентативній вибірці з цільової популяції осіб.
8. Визначення за даними апробації з попереднього етапу статистичних характеристик тестових завдань, відбір до тесту завдань з задовільними характеристиками.
9. Дослідження надійності та валідності для кінцевого набору тестових завдань.
10. Розробка інструкції для адміністрування (процесу пред'явлення) тесту, визначення шкали вимірювання, інтерпретація окремих положень екзаменованих на шкалі, визначення норм, порогових оцінок тощо.

Перелічені етапи складають мінімальний набір дій, які можуть за потреби повторюватися та містити додаткові дії. У цій главі розглянемо перші шість етапів, змісту решти чотирьох присвячені наступні глави.

1. Визначення мети вимірювання. Без перебільшення можна стверджувати, що без правильного позиціювання мети вимірювання тестування приречене на невдачу. Загальною метою усіх вимірювань є прийняття тих чи інших рішень. Прикладами рішень є: рішення про внесення змін до змісту та методів навчання (діагностичне вимірювання), рішення про відповідність рівня навчальних досягнень учня з певної дисципліни заданим критеріям (тестування рівня навчальних досягнень), рішення про прийом вступника до ВНЗ чи віднесення студента до однієї з навчальних груп (відбіркове тестування), рекомендації щодо вибору особою майбутньої професії (професійна орієнтація), рішення щодо прийому особи на вакантну посаду (професійний відбір).

В навчальних закладах часто практикуються тести поточного чи підсумкового контролю та діагностичні тести. Зазвичай конструювання подібних тестів відбувається за скороченою процедурою, викладачі розробляють тестові завдання, покладаючись переважно на власний досвід та інтуїцію. Такі тести не претендують на роль справжнього вимірювання хоча б тому, що для них не з'ясовуються необхідні психометричні характеристики шляхом апробації на репрезентативних вибірках суб'єктів вимірювання. Це не означає, що ці тести не корисні, якщо вони гармонійно вписані у цілісну систему оцінювання навчальних досягнень учнів чи студентів. Часом, щоб підкреслити відмінність цих тестів від педагогічного вимірювання, їх називають контрольними роботами (іспитами, завданнями) у тестовій формі.

На роль вимірювання зазвичай можуть претендувати програми тестування у регіональному чи національному масштабах. Прикладом такої програми є зовнішнє незалежне оцінювання випускників загальноосвітніх шкіл в Україні. Становлення ЗНО супроводжувалось певними відхиленнями від стандартних процедур конструювання тесту, пов'язаними з особливостями освітньої політики керівних освітянських органів та, часто, браком коштів. Зокрема, протягом певного часу тести ЗНО проводилися з подвійною метою: для визначення навчальних досягнень випускників з окремих шкільних предметів та для відбору до ВНЗ. Очевидно, що і та, і інша мета мають свою специфіку і вимірювання для досягнення кожної з них має бути, взагалі кажучи, різним. При оцінюванні рівня навчальних досягнень з певної дисципліни головну

роль відіграє з'ясування відповідності знань та умінь учнів програмі шкільного курсу. Інша справа – відбір для навчання у ВНЗ. Тут на перший план виходять загальні навчальні компетенції, здатність учня до навчання за тим чи іншим професійним напрямком. Тому для оптимального відбору бажано поєднувати кілька предметних тестів з тестом загальної навчальної компетентності, і, можливо, додатковими критеріями, такими, як середній бал атестата випускника школи. У США, де тестування дуже поширене, більшість університетів вимагають від вступників проходження тесту SAT I, який є тестом на здібність до навчання, а також, за вибором університету, тих чи інших предметних тестів, які об'єднані під загальною назвою SAT II.

Одним із важливих рішень, яке повинен прийняти розробник тесту, і яке сильно залежить від мети вимірювання, є рішення про те, чи повинен тест бути критеріально-орієнтованим, чи нормо-орієнтованим. Тести на професійну придатність, очевидно, повинні бути критеріально-орієнтованими. Тести здібностей можуть бути нормо-орієнтованими. Для останніх важливо забезпечити потрібну роздільну здатність тесту шляхом включення до нього достатньо великої кількості завдань середньої складності.

2. Визначення репрезентативних поведінкових характеристик. Фахівці одностайно вказують на неформальність цього етапу конструювання тесту. Він вимагає від розробника значного досвіду та інтуїції. Разом з тим, існує ряд методів визначення оптимальних для вимірювання поведінкових характеристик суб'єктів вимірювання (операціоналізації конструкту), з яких той чи інший метод може підходити в залежності від того, яка риса чи конструкт вимірюється. Крокер та Алгіна [6] наводять наступний перелік таких методів.

1. Аналіз змісту (контент-аналіз). У суто психологічному тестуванні, яке, проте, може становити значний інтерес і у освітньому аспекті, для визначення основних проявів вимірюваного конструкту можуть аналізуватися тексти учнів (твори, есе) на задану тему. Цей метод використовувався, наприклад, для конструювання популярних опитувальників, призначених для вимірювання здатності дітей до самоконцептуалізації, Я-концепції, самооцінки.

2. Огляд досліджень. Розробник тесту повинен бути обізнаним з науковими результатами, отриманими раніше різними фахівцями щодо визначення конструкту, який вимірюється. Ці результати потрібно використовувати або брати до уваги у тій мірі, у якій розробник вважає за доцільне.

3. Критичні ситуації. Конструкт представляється зазвичай у вигляді континуума на прямій. З огляду на це важливо знайти екстремальні точки цього континууму. Наприклад, для оцінювання результатів роботи службовців розробник тесту може запропонувати керівникам описати ситуації, у яких їх підлеглі працювали з найбільшою (чи найменшою) ефективністю.

4. Прямі спостереження. У цьому випадку розробник організовує особисте спостереження за поведінкою суб'єктів вимірювання. Наприклад, для розробки тесту на оцінювання стресу, викликаного роботою в небезпечних ситуаціях, розробник може самостійно фіксувати рівень стресу працівників у різних робочих ситуаціях. Таким чином можна визначити, які саме ситуації пов'язані з різними рівнями стресу.

5. Експертні судження. Розробник може звернутися за допомогою до експертів з даної області. Інформацію від експертів зазвичай отримують у вигляді відповідей на спеціальні анкети або індивідуальних інтерв'ю.

6. Навчальні цілі. Цей метод становить особливий інтерес для розробників тестів навчальних досягнень. Програми навчальних курсів зазвичай містять перелік знань і умінь, які повинні продемонструвати учні чи студенти після завершення курсу. У вищих навчальних закладах переліки знань та умінь, визначені програмами курсів, повинні узгоджуватися з переліком компетенцій, наведених у освітньо-кваліфікаціних характеристиках (ОКХ) даної спеціальності. Ті, в свою чергу, повинні вписуватися у національні рамки кваліфікацій та більш широкі рамки, такі, як рамка кваліфікацій Європейського простору вищої освіти. Крім того, ОКХ спеціальності як частина стандарту вищої освіти для даної спеціальності містить додаткову інформацію, яка вказує на певні поведінкові характеристики осіб стосовно визначених знань та умінь: види типових задач діяльності (професійна, соціально-виробнича, соціально-побутова), класи задач діяльності (стереотипна, евристична, діагностична), види уміння (предметно-практичне, предме-

тно-розумове, знаково-практичне, знаково-розумове), рівні сформованості уміння (здатність виконувати дію, спираючись на матеріальні носії інформації щодо неї; здатність виконувати дію, спираючись на постійний розумовий контроль без допомоги матеріальних носіїв інформації; здатність виконувати дію автоматично, на рівні навички).

В ідеалі, розробка переліку знань та умінь з даного навчального курсу повинна супроводжуватися одночасною розробкою діагностичних засобів, з урахуванням вимог ОКХ. Основні діагностичні засоби повинні бути складовою освітнього стандарту.

Іншими важливими джерелами інформації для розробника тесту є реальні навчальні матеріали, які пред'являлися учням (студентам), а також дані про рівні засвоєння необхідних знань та умінь конкретними особами (особливо для нормо-орієнтованого тестування).

Важливо розуміти, що предметна область і область тестування практично ніколи не збігаються. Тестування як формалізована процедура накладає ряд обмежень на можливість спостерігати прояви риси чи конструкту у екзаменованих у різноманітних умовах та ситуаціях. Більше того, з метою забезпечення рівних умов тестування екзаменовані під час пред'явлення тесту ставляться у однакову для всіх ситуацію. Вибір способу тестування (наприклад, паперове чи комп'ютерне) теж накладає обмеження на можливі прояви вимірюваної якості. Наприклад, при письмовому тестуванні з іноземної мови без залучення додаткових технічних засобів чи фахівців – носіїв мови неможливо здійснити аудіювання. Таким чином, область тестування завжди є вужчою у порівнянні з предметною областю. Процес визначення області тестування називають вибіркою змісту (domain sampling). У широкому розумінні область тестування включає у себе необмежену популяцію можливих тестових завдань. У більш вузькому розумінні – це набір специфікацій тестових завдань, який дозволяє конструювати практично будь-яку кількість паралельних (однакових за змістовими та статистичними характеристиками) форм тесту. Створені в межах однієї специфікації завдання вважаються еквівалентними і можуть заміняти одне одного. Специфікації тестових завдань можуть включати інформацію про зміст завдань, описання проблем-

них ситуацій чи стимулів, характеристики правильних відповідей, характеристики дистракторів для завдань множинного вибору.

З огляду на важливість цього поняття наведемо повністю приклад специфікації тестового завдання з книги Крокер та Алгіни [6 , с. 110]. У цьому прикладі передбачається конструювання тесту з базового рівня математики (таблиця 3.1).

Таблиця 3.1. Специфікація тестового завдання з математики

Ознаки стимулу	**Ознаки відповіді**
1. Завдання повинне пропонувати помножити два десяткові числа, дробові і/або змішані. 2. Завдання повинне бути сформульованим або у вигляді тексту, або у вигляді горизонтального рядку чисел. Якщо використовується горизонтальний рядок, та інструкція може мати вигляд «Помножте» або «Знайдіть добуток». 3. Один із множників повинен містити три цифри, відмінні від нуля; інший множник повинен також містити три цифри, дві з яких повинні бути більшими від п'ятірки. 4. Кожен множник повинен мати принаймні один десятковий знак після коми. 5. Добуток повинен містити не більше чотирьох десяткових знаків після коми. 6. Розв'язання повинне містити перенесення коми щонайменше на дві цифри. 7. При виборі множників жодна цифра не повинна використовуватися більше двох разів.	1. Формат: усі числові варіанти відповідей повинні мати вигляд десяткових дробів і розташовані у порядку зростання або спадання чисел у місці відповіді. 2. Завдання повинне містити чотири варіанти відповіді: а) правильна відповідь; б) неправильна відповідь, яка відображає помилку у перегрупуванні цифр після коми у результаті множення; в) неправильна відповідь, яка відображає помилку в цифрах у результаті множення; г) неправильна відповідь, яка відображає або пропущення коми у десятковому дробі, або помилку у розташуванні коми у десятковому дробі. 3. Інша можлива відповідь – «Жодне з переліченого» не може заміняти варіанти а)-в). Цей вибір може використовуватися лише замість четвертого варіанту.

Ще більш структуровану специфікацію тестового завдання можна отримати, використовуючи алгоритм побудови завдання або фасет завдань. Алгоритм особливо добре підходить для завдань у вигляді формул. Прикладом використання алгоритму є один із видів завдань модуля тестування у популярній інтернет-системі управління навчанням Moodle. Фасети використовуються для тестових завдань з вербальним змістом. Фасет має фіксовану синтаксичну структуру з одним або кількома змінними елементами. Наприклад, це може бути речення з пропущеними словами, до кожного з яких додається набір можливих значень. Лише певні визначені поєднання цих значень у реченні є правильними відповідями.

Існують й інші методи створення специфікацій тестових завдань. Зауважимо, що використання специфікацій накладає особливу відповідальність на їх розробника.

Особливо корисно використовувати створення специфікацій завдань у ситуаціях, коли у розробленні тесту беруть кілька авторів. Якщо кожен автор буде користуватися специфікаціями тестових завдань, то за короткий період часу може бути створено кілька паралельних форм тесту.

3. Підготовка специфікації тесту. Після виконання попередніх етапів розробки тесту автор повинен скласти план прийняття рішень відносно пропорцій кожного з компонентів вимірюваної якості, які будуть представлені у тесті. У більшості випадків створення тесту навчальних досягнень автор повинен керуватися двома чинниками: змістовим наповненням завдань та рівням пізнавальної (когнітивної) діяльності, які повинен продемонструвати екзаменований. Останній може відповідати одній із відомих таксономій, наприклад, таксономії Блума (знання – розуміння – застосування – аналіз – синтез – оцінювання).

У випадку врахування обох чинників специфікація тесту може мати вигляд матриці – таблиці, у якій вздовж однієї сторони розміщуються змістові компоненти предметної області, вздовж іншої – необхідні рівні пізнавальної діяльності. Розглянемо як приклад специфікацію тесту для атестаційного екзамену на звання вчителя американського штату Флорида (Крокер, Алгіна, с. 116, перелік номерів компетенцій опустимо).

Таблиця 3.2. Приклад специфікації тесту

Змістові базові категорії	**Рівень**			**Всього**
	Знання	**Застосування**	**Вирішення проблем**	
1. Керування класом (компетенції 6, 12, 13, 15, 16, 17, 20, 22)	1	10	9	20
2. Розвиток учнів (компетенції …)	1	11	11	23
3. Оцінювання, реєстрація, повідомлення про прогрес учня (компетенції …)	2	16	1	19
4. Навчальні матеріали (компетенції …)	1	6	2	9
5. Навчальні цілі (компетенції …)		7	3	10
6. Навчання і научання (компетенції …)	2	12	5	19
Всього	7	62	31	100

У інших випадках специфікація тесту може містити лише перелік кількості завдань відповідно до змістових компонентів предметної області, або додатково визначати, для яких компонентів які й скільки видів тестових завдань буде використовуватися,

якої трудності будуть завдання. Для прикладу розглянемо специфікацію тесту ЗНО з математики 2010 року (матеріал узято з сайту Сімферопольського РЦОЯО).

Специфікація містить інформацію про призначення тесту, документи, що визначають зміст тесту з математики, загальну кількість завдань тесту, скільки часу відведено на виконання тесту, характеристику типів тестових завдань, принцип розташування завдань у тестовому зошиті, правила оцінювання завдань кожного з типів, додаткові матеріали та обладнання (не передбачалися), умови проведення зовнішнього незалежного оцінювання та перевірки його результатів, рекомендації з підготовки до складання тесту.

Крім того, специфікація містить три таблиці. У кожній з них вздовж вертикальної сторони наведений перелік змістових компонентів. Вздовж горизонтальної сторони наведено розподіл кількості завдань:

- у першій таблиці – за типами завдань (див. таблицю 3.3);
- у другій – за складністю (таблиця 3.4);
- у третій – за рівнями когнітивної діяльності (таблиця 3.5).

Таблиця 3.3. Розподіл кількості завдань за змістом і типами

Навчальний предмет	**Змістові лінії**	**Кількість завдань**			**%**
		Завдання з вибором правильної відповіді	**Завдання на встановлення відповідності**	**Завдання з короткою відповіддю**	
Алгебра і початки аналізу	Числа і вирази	6	1	1	22,2
	Рівняння і нерівності	1	-	3	16,7
	Функції	3	1	2	16,7
	Елементи комбінаторики, початки теорії ймовірностей та елементи статистики	3	-	-	8,3
Геометрія	Планіметрія	5	-	1	16,7
	Стереометрія	5	1	1	19,4
Усього		25	3	8	100

Таблиця 3.4. Розподіл кількості завдань за змістом і складністю

Навчальний предмет	**Змістові лінії**	**Складність завдань**			**Разом**
		Легкі (0,9-0,7)	**Оптимальні** (0,6-0,4)	**Складні** (0,3-0,1)	
Алгебра і початки аналізу	Числа і вирази	2	6	-	8
	Рівняння і нерівності	1	2	3	6
	Функції	1	3	2	6
	Елементи комбінаторики, початки теорії ймовірностей та елементи статистики	-	3	-	3
Геометрія	Планіметрія	2	3	1	6
	Стереометрія	1	5	1	7
Усього		7	22	7	36

Таблиця 3.5. Розподіл кількості завдань за змістом і рівнями пізнавальної діяльності

Навчальний предмет	**Види діяльності** / **Змістові лінії**	**Знання і розуміння**			**Застосування знань і умінь в типових та змінених ситуаціях**			**Застосування знань і вмінь у нових ситуаціях**		
		ВОВ	ВВ	КВ	ВОВ	ВВ	КВ	ВОВ	ВВ	КВ
Алгебра і початки аналізу	Числа і вирази	2	1	-	3	-	1	1	-	-
	Рівняння і нерівності	-	-	-	3	-	1	-	-	2
	Функції	1	1	-	2	-	1	-	-	1
	Елементи комбінаторики, початки теорії ймовірностей та елементи статистики	1	-	-	2	-	-	-	-	-
Геометрія	Планіметрія	1	-	-	3	-	-	1	-	1
	Стереометрія	1	1	-	4	-	1	-	-	-
Усього (%)		17	8	-	47	-	11	6	-	11

4. Конструювання завдань. Перш за все, розробник тесту має визначитися з типами тестових завдань, якщо ця робота не була виконана на попередніх етапах конструювання тесту. Після того, як типи завдань визначені, автори завдань повинні вивчити наявну літературу з методики створення таких завдань. Зауважимо, що початківці часто переоцінюють свої здібності у цьому виді роботи.

Існує багато типів завдань, придатних для тесту навчальних досягнень. Найбільш часто використовуються три типи об'єктивних завдань (або, як у нас їх частіше називають, завдань закритого типу). Цими популярними типами завдань є

1) завдання з альтернативним вибором;
2) завдання з множинним вибором;
3) завдання на відповідність.

Важливою ознакою якості всіх трьох типів завдань є правдоподібність неправильних варіантів відповіді. Часто, як ми це бачили у наведеному раніше прикладі специфікації завдання множинного вибору на множення двох чисел, дистрактори відображають поширені помилки у розв'язанні завдань. Для цього автори повинні бути добре знайомими з популяцією екзаменованих.

Цікаво відмітити, що можна сконструювати завдання, яке б у іншій формі не мало б сенсу. Приклад завдання множинного вибору з однією правильною відповіддю:

Господиня придбала яйця у супермаркеті. Третина яєць розбилася. Скільки яєць було придбано?

а) 7
б) 8
в) 9
г) 10

Протягом усієї історії тестування ведуться дискусії між прихильниками та противниками використання завдань множинного вибору. Противники зазвичай вказують на істотну ймовірність вгадування екзаменованим правильної відповіді. Прихильники відповідають, що протягом достатньо короткого часу екзаменованому можна пред'явити велику кількість таких завдань, що, по

суті, нівелює ефект угадування, покриваючи разом з тим велику частину предметної області. Противники вказують також на нібито неможливість перевіряти за допомогою таких завдань високі рівні когнітивної діяльності. У цьому аспекті показовим є намагання зіставити завдання множинного вибору з завданнями відкритого типу особливого виду – з короткою відповіддю (йдеться про завдання, яке вимагає дати відповідь у вигляді одного-двох слів, наприклад, речення з пропущеним словом; такі відповіді може перевіряти й комп'ютер). Автори-початківці зазвичай схиляються до використання останніх, аргументуючи свій вибір тим, що ймовірність угадування відповіді на таке завдання практично дорівнює нулю. Насправді завдання з короткою відповіддю мають дуже важливий недолік: за їх допомогою загалом важко перевіряти високі рівні когнітивної діяльності екзаменованих. Завдання ж множинного вибору зазвичай недооцінюються у цьому аспекті. Зауважимо, що у цьому питанні не останню роль відіграють особливості предметної області. У тесті з математики завдання, яке передбачає відповідь у вигляді одного числа, може перевіряти фактично будь-які рівні пізнавальної діяльності, і його легко реалізувати як у одній, так і у іншій формі. У гуманітарних дисциплінах завдання з короткою відповіддю загалом значно програють завданням множинного вибору у аспекті визначення рівня когнітивної діяльності.

Схоже на те, що представників гуманітарних наук бентежить деяка механічність, яка супроводжує тестування з завданнями множинного вибору. Так само їх бентежило б, якби вони до цього вже не звикли, той факт, що будь-який цифровий мультимедійний продукт, наприклад, високохудожня музика чи фільм, кодується лише одними нулями та одиницями – аналогами відповідей на найпростіші завдання з альтернативним вибором. Цей факт, до речі, ілюструє крайню, ультимативну точку зору прихильників завдань множинного вибору, яка полягає у тому, що за їх допомогою можна перевіряти все що завгодно. Але це потрібно уміти робити! Написати геніальну музику може тільки геній, закодувати її найпростішими засобами – справа технології. Автори тестових завдань повинні достатньо добре робити і те й інше. Достатньо часто від людей, не обізнаних з експериментальною психологією та психометрією, можна почути, що тестуванням взагалі неможливо перевіряти вираженість таких конструктів, як, скажімо, креати-

вність. Корінь проблеми тут криється у небажанні детально операціоналізувати конструкт. Латентна характеристика особи може проявлятися у різних побічних поведінкових реакціях на стимул. Ситуація схожа з тією, коли справжню кількість населення Києва визначають за кількістю проданого хліба. Прикладом завдання множинного вибору на креативність може бути наступне:

Ви маєте намір відвідати країну, в якій ще не бували. Якому варіанту подорожі ви віддасте перевагу?

а) Стандартному туру з оглядом визначних пам'яток під керівництвом гіда.

б) Організованому туру з наявністю достатнього вільного часу.

в) Самостійній подорожі за власним планом.

Звичайно ж, одне таке завдання не дасть змоги виявити рівень креативності людини. Але достатньо велика кількість подібних завдань дозволяє це зробити. Заради справедливості потрібно зазначити, що тести на креативність зазвичай містять більш хитромудрі форми завдань – наприклад, домалювати картинку, чи придумати максимально багато способів використання якогось предмету. Але такі завдання вимагають високої кваліфікації як автора, так і особи, яка буде перевіряти відповіді. Чи не є це підставою вважати, що тест з подібними завданнями – це не тест взагалі? Прискіпливий читач уже мабуть помітив, що ми уникаємо строгого визначення поняття тесту. Ми використовуємо цей термін як синонім інструменту вимірювання. В свою чергу, інструмент може вважатися вимірювальним, якщо визначені такі його характеристики, як валідність, надійність, шкала вимірювання, роздільна здатність, стандартна похибка вимірювання. З цієї точки зору можна навіть перетворити на тест, наприклад, фіксацію дій учня під час виконання ним лабораторної роботи з фізики чи з інформатики.

З іншого боку, необхідно враховувати зворотний ефект, який тестування чинить на навчальний процес. Скажімо, використання завдань з розгорнутими відповідями спонукає учня до вдосконалення уміння висловлювати свої думки засобами мови, а оцінювачу надає можливість побачити це уміння, безпосередньо спостерігати за логікою дій екзаменованого.

В Україні зараз переважає компромісний підхід щодо вибору типів завдань для тестування. Вважається, що тести повинні містити невелику кількість есе – завдань, що вимагають розгорнутої відповіді. Для того, щоб оцінювання есе було максимально об'єктивним, розробники тесту надають оцінювачам детальну інструкцію з критеріями оцінювання. Для прикладу наведемо критерії оцінювання на оцінки 4 та 3 за 6-бальною шкалою, які використовувалися у тесті здібностей SAT (за матеріалами буклету з підготовки до тесту у 2008 році, перекладено журналом Вісник ТІМО, №5 за 2009 р.):

Оцінка 4	Оцінка 3
Есе у цій категорії демонструє достатній рівень майстерності, хоча має певні огріхи в якості. Типове есе:	Есе у цій категорії демонструє рівень майстерності, що вдосконалюється, і має один або декілька з наступних недоліків
• розвиває точку зору на певне питання і демонструє компетентне критичне мислення з використанням достатніх прикладів, доведень і інших доказів на підтримку позиції автора	• розвиває точку зору на певне питання і демонструє деяке критичне мислення, проте може це робити не послідовно з використанням невідповідних прикладів, доведень або інших доказів на підтримку позиції автора
• як правило, організоване і зосереджене, демонструючи певне об'єднання і розвиток ідей	• обмежене своєю організованістю чи фокусуванням, або може демонструвати неточності в об'єднанні чи розвитку ідей
• представляє відповідне проте непослідовне використання мови, зазвичай застосовуючи відповідний словниковий запас	• представляє вміння використання мови, що розвивається, проте інколи використовує недостатній словниковий запас або невідповідний вибір слів
• демонструє деяку різноманітність структур речень	• демонструє недостатню різноманітність або проблеми у структурах речень
• має деякі граматичні і механічні помилки	• містить цілу низку граматичних і механічних помилок

Для підвищення об'єктивності оцінювання есе організатори тестування намагаються залучати до їх перевірки принаймні двох незалежних оцінювачів, а у випадку протиріч між тими з приводу якоїсь оцінки – третього оцінювача.

Для опитувальників найбільш характерними формами завдань є завдання альтернативного вибору (зазвичай з варіантами відповіді «погоджуюся – не погоджуюся», завдання у формі Лайкерта (Likert), і біполярний (двосторонній) список.

Завдання альтернативної форми містить деяке твердження, з яким опитуваному пропонується погодитися або не погодитися. Важливо розуміти, що тут немає правильних або неправильних відповідей. Якщо передбачається, що більша кількість балів за тест повинна відповідати більшій вираженості у опитуваного конструкту, для вимірювання якого складено тест, то вищий бал за завдання повинен ставитися за ту відповідь, яка відповідає більшій вираженості конструкту. Це може бути відповідь як «погоджуюся», так і «не погоджуюся». Наприклад, відповідь «погоджуюся» на завдання з твердженням «Діти повинні беззаперечно підкорятися своїм батькам» свідчить про більшу вираженість атитюду, який може бути описаним як континуум від авторитарної до поблажливої точки зору ставлення до батьківської дисципліни, у бік авторитарної. Якщо найбільша кількість балів передбачається для особи з найбільш поблажливою точкою зору, то за цю відповідь може присвоюватися 0 балів, а за відповідь «не погоджуюся» – 1 бал. Існує також підхід, який передбачає зважування завдань з огляду на те, наскільки сильно виражено вимірюваний конструкт у твердженні даного завдання. Терстоуном описана методика, за якою автор опитувальника розробляє велику кількість тверджень (Терстоун пропонував 100), які розташовуються у порядку від вкрай позитивного ставлення до конструкту до вкрай негативного. Деякі твердження мають бути нейтральними щодо ставлення до конструкту. Всі твердження поміщаються групою експертів у один з 7, 9, або 11 рівних за довжиною інтервалів, на які розбито континуум конструкту. Після цього ті твердження, думки експертів щодо яких найбільше розходилися, відсіваються, а тим, що залишаються, присвоюють вагові коефіцієнти, які відповідають інтервалам, у яких вони знаходяться. Коли опитуваний погоджується з твер-

дженням, його ваговий коефіцієнт додається до загальної оцінки опитуваного.

Інша широко розповсюджена форма завдань для опитувальників запропонована Лайкертом у 1932 році. Завдання за цією формою містить твердження, яке виражає вкрай позитивне або вкрай негативне відношення до конструкту. Далі наводиться п'ять варіантів відповіді: «повністю не погоджуюся» - «не погоджуюся» - «не упевнений» - «погоджуюся» - «повністю не погоджуюся». При оцінюванні відповіді на таке завдання 1 бал присвоюється за відповідь, яка вказує на найбільш низьку вираженість конструкту, 2 бали – за наступну і т.д. Загальною оцінкою за тест буде сума балів, отриманих за кожне з завдань.

Ще одна популярна форма завдань для опитувальників – біполярна пара (пара антонімів). Приклад:

Дитина з затримкою у розвитку
Симпатична - - - - - Потворна
Радісна - - - - - Сумна
Брудна - - - - - Чиста

У цьому завданні між антонімами кожної пари помічено п'ять точок континууму. Для кожної пари опитуваний повинен вибрати точку, яка найбільш точно відповідає його почуттю.

Більшість пар, за Осгудом, можуть бути згруповані в одному з трьох вимірів (фактор оцінки, фактор сили, фактор активності). Метод аналізу та інтерпретації відповідей на такі завдання отримав назву *семантичного диференціалу*.

Розробник опитувальника повинен пам'ятати про феномен, який називають *схильністю до відповідей*. Найбільш крайні прояви цієї схильності опитуваних можуть сильно впливати на правильність інтерпретації відповідей – поступливість (схильність погоджуватися з твердженнями незалежно від їх змісту), та диференційована індивідуальна інтерпретація невизначених специфікаторів, таких як *інколи* і *часто*.

5. Рецензування. В організаціях, які займаються виготовленням стандартизованих тестів (часто це незалежні комерційні компанії) дуже відповідально ставляться до етапу рецензування новостворених тестових завдань. Група рецензентів зазвичай повинна включати експертів з фаху, до якого відноситься вимірюва-

на якість, експертів з вимірювань, а також фахівців з мови. Ключові напрямки рецензування:

- точність і однозначність формулювання завдання;
- доречність, відповідність специфікаціям;
- технічні недоліки, пов'язані з конструкцією завдання;
- граматика та орфографія;
- неупередженість (однакове сприйняття завдання представниками різних соціальних чи культурних груп у популяції екзаменованих;
- зручність прочитання.

Фахівці з цільової області вимірювання повинні переконатися, що завдання тесту не допускає неоднозначної відповіді. Особливо це стосується завдань у закритій формі. Також слід упевнитися, що зміст завдання дійсно відповідає, з одного боку, меті вимірювання, з іншого боку – вимірюваній якості, а також цільовій популяції. Цим самим буде забезпечена так звана *змістова валідність* завдання. У деяких випадках *тестування високої відповідальності* (*high stake testing*), тобто тестування, результати якого можуть кардинально вплинути на життєвий шлях екзаменованого, наприклад випускний тест або тест для вступу до університету, бажано забезпечити також *очевидну валідність* (*face validity*) – зрозумілість і прийнятність завдання для нефахівців, які можуть проте бути зацікавленими особами (наприклад, батьків чи політиків). Якщо завдання створювалося на основі специфікації, то потрібно перевірити відповідність завдання специфікації.

Тестові завдання можуть володіти недоліками, притаманними для різних форм завдань. Існують недоліки, загальні для форми завдання, наприклад, коли у завданні множинного вибору варіант з правильною відповіддю значно довший у порівнянні з дистракторами. Разом з тим, існують недоліки, специфічні для цільової області вимірювання. Наприклад, математики добре знають, що не можна у завданні «розв'язати рівняння…», правильною відповіддю на яке є одне число, формулювати варіанти відповідей у вигляді чисел, оскільки екзаменований може знайти правильну відповідь, просто підставляючи кожен з варіантів у рівняння. У даному випадку проблему легко усунути, змінивши форму завдання на

відкриту. Підкреслимо, що з правилами побудови завдань у різних формах повинні бути обізнані як автори, так і рецензенти.

Фахівці з мови повинні перевірити відсутність у завданні орфографічних та граматичних помилок, а також лаконічність тексту.

Ще один важливий аспект – забезпечення справедливості щодо різних соціальних та/або культурних груп, представники яких можуть зустрічатися у цільовій популяції. Хрестоматійним є приклад, який свого часу набув широкого розголосу у США. У одному з завдань тесту SAT екзаменованим пропонувалося підібрати пару слів, найбільш близьких за логічним зв'язками до пари *бігун-марафон*. Правильною відповіддю була пара *весляр-регата*. На це завдання дали правильну відповідь 53% білих учнів, і лише 22% чорношкірих. Політики звинуватили розробників SAT в упередженому ставленні до бідних чорношкірих представників населення у порівнянні з білими багатими власниками яхт, оскільки перші могли не знати і ніколи не чути слова *регата*.

Якщо метою тестування не є перевірка швидкості читання і сприйняття тексту, потрібно перевірити, чи не буде довжина тексту завдання впливати на правильність відповіді. Особливо це важливо для цільових популяцій, для членів яких мова тестування не завжди є рідною, а також у випадку тестування дітей дошкільного або раннього шкільного віку. Надто довге формулювання завдання і/або відповідей може погіршувати змістову валідність навіть для однорідної за усіма ознаками популяції.

У випадку обмежених матеріальних ресурсів рецензування завдань можна здійснювати після їх попередньої апробації на невеликій вибірці представників популяції екзаменованих, яка дозволить частину завдань відсіяти або вкаже на необхідність їх коригування.

6. Попередня апробація завдань. Цей етап є необов'язковим, але бажаним перед широкомасштабними (польовими) випробовуваннями тестових завдань на репрезентативній вибірці з цільової популяції, оскільки дозволяє зекономити ресурси шляхом вилучення чи коригування завдань. Завдання, призначені для тестування високої відповідальності, бажано попередньо перевірити на вибірці 100-200 осіб, а якщо такої можливості немає, то хоча б на вибірці об'ємом не менше 15 осіб. Під час попередньої

апробації розробники безпосередньо спостерігають за виконанням завдань, фіксуючи моменти, які викликають у учасників тестування збентеження. Після сеансу тестування з його учасниками проводять співбесіду, під час якої просять їх прокоментувати завдання і запропонувати шляхи їх вдосконалення.

Рекомендується також зібрати та проаналізувати описові статистичні характеристики завдань, для того щоб мати приблизні уявлення про їх складність, дисперсію відповідей, диференційовну здатність. Все це також дозволить скоригувати деякі завдання ще до початку польового випробовування.

Наступні етапи. Після проведення попередньої апробації завдань проводиться польове випробування їх на репрезентативній вибірці з цільової популяції у вигляді та умовах, максимально наближеними до тих, які передбачаються при повноцінному (можливо, й комерційному) використанні тесту. Цей етап є надзвичайно важливим, оскільки саме на його основі отримують психометричні характеристики завдань, проводять дослідження валідності та надійності, і, врешті, формують остаточний варіант тесту з заданими характеристиками. Якщо передбачається нормо-орієнтоване використання тесту, то завдання для кінцевого набору обирають, виходячи з статистичних характеристик завдань. Для нормо-орієнтованого тестування також існують методики вибору завдань.

Для остаточно створеного тесту проводять дослідження надійності та валідності, які характеризують тест в цілому.

Усі ці дії будуть розглядатися нами в наступних розділах.

4. ВАЛІДНІСТЬ: ЗАГАЛЬНИЙ ОГЛЯД

Поняття валідності та валідизації інструментів вимірювання є одним з чи не найскладніших для розуміння та застосування в теорії освітніх вимірювань. Необхідність валідизації нових тестів є наслідком самої природи тестування, яка передбачає, з одного боку, наявність певних попередніх теоретичних конструкцій, а з іншого боку, – відшукання часткових свідчень правдоподібності цих конструкцій, подальшого узагальнення свідчень, а також правильної інтерпретації та застосування результатів тестування.

У змісті поняття валідності, починаючи з 20-х років минулого століття, акцент поступово зміщувався від концентрації навколо ідеї відповідності результатів тестування зовнішньому критерію (критеріальна валідність), до необхідності оцінювання насамперед *правильності інтерпретації та використання* результатів тестування. На цьому шляху теорія валідизації пройшла через період панування так званої «троїстої» моделі валідності (змістова, критеріальна та конструктна типи валідності), до виділення конструктної валідності як загальної, що охоплює також поняття змістової та критеріальної валідності, і далі – до трактування валідності як аргументу в рамках загальної теорії валідизації М. Кейна.

Модель Кейна валідності як аргументу запропонована ним ще у 1982 році, але тоді ця теорія не набула помітного поширення. Однак саме Кейну було довірено написати розділ «Валідизація» для четвертого видання фундаментального збірника «Educational Measurement» [10], який видано у 2006 році під егідою ACE та NCME. Підхід Кейна є спробою надати розробникам та користувачам тестів єдину методологію валідизації, і це є добрим аргументом для покладання цієї методології в основу загального дослідження валідності тесту. Далі ми намагатимемося прослідкувати логіку розвитку поняття валідності та валідизації стосовно інструментів освітніх вимірювань, стисло викласти сучасний підхід до цієї проблеми, узагальнений М. Кейном, а також розглянути основні конкретні методи дослідження того чи іншого виду валідності.

Еволюція поняття валідності в тестуванні. Валідність в науці – це відповідність емпіричних досліджень тій меті, заради якої вони проводяться.

В психометрії тест називається *валідним*, якщо він адекватно вимірює саме ту якість (рису чи конструкт), для вимірювання якої він був створений.

Укладачі та користувачі тестів часто обмежуються лише перевіркою часткових свідчень валідності. У деяких випадках це цілком виправданий підхід. Наприклад, якщо вчитель в кінці уроку пропонує учням невеликий тест на засвоєння щойно пройденого матеріалу, достатньо, крім визначення правила нарахування балів за окремі завдання та правильної процедури проведення тестування, забезпечити також змістову валідність цього тесту. В інших випадках, за наявності валідного критерію (наприклад, іншого тесту, який використовувався з тією ж метою), цілком достатньо перевірити, «за рівних інших умов», чи достатньо високою є кореляція результатів тестування з результатами застосування критерію. Але для виготовлення стандартизованого інструменту вимірювання, для якого передбачається масове повторне використання, і результати якого можуть інтерпретуватися й застосовуватися у різних контекстах, дослідження валідності має бути комплексним, тобто воно повинне охоплювати усі її аспекти.

Для кращого розуміння проблеми валідності розглянемо значно простіший, з точки зору валідизації, вид вимірювання – фізичне вимірювання, наприклад, вимірювання температури тіла людини. Очевидно, що, перш за все, нам потрібен якісний інструмент – градусник. Чи можемо ми замість градусника використати кімнатний термометр? Якщо так, то з якими відмінностями й обмеженнями? Якість градусника можна визначати по-різному. Наприклад, ми можемо порівняти його показання з показаннями іншого градусника, якість якого не викликає у нас сумнівів («критеріальна» валідність), або перевірити показання нашого градусника на достатньо великій групі цілком здорових осіб. Крім того, ми повинні правильно «читати» шкалу градусника. Чи виконане градуювання шкали за Цельсієм, чи за Фаренгейтом? Нехай ми упевнилися у добрій якості градусника і вміємо зчитувати його показання. Чи є це достатніми для того, щоб отримати правильні результати його конкретного використання? Очевидно, ні, оскільки

неправильною могла бути процедура вимірювання, наприклад, замість необхідних семи хвилин, пацієнт тримав градусник дві хвилини, або тримав градусник не під пахвою, а у кишені пальто. Якщо градусник має добру якість і процедура вимірювання була правильною, то виникає наступна проблема – як слід інтерпретувати отриманий результат? Що означає, наприклад, що градусник показує температуру 37 градусів за Цельсієм? Яка температура вважається нормальною, і чи є ця норма застосовною також і для конкретного пацієнта? Зауважимо, що інтерпретація отриманого результату вимірювання температури тіла залежить також і від мети, з якою проводилося вимірювання. Наприклад, якщо лікар хоче перевірити, як подіяв на пацієнта, у якого напередодні була висока температура, призначений йому лікувальний засіб, він може інтерпретувати температуру в 37 градусів як позитивний результат. Навпаки, якщо пацієнт звернувся до лікаря вперше, це ж саме показання повинне трактуватися лікарем як «підвищена температура». У цьому випадку лікар може інтерпретувати показання градусника як симптом можливої хвороби. Хвороба на цій початковій стадії трактується лікарем як комплекс відповідних симптомів, серед яких температура тіла відіграє певну визначену наперед роль: для одних хвороб це важливий показник, для інших – менш важливий, або й зовсім не важливий. Нарешті, на основі отриманого результату вимірювання лікар повинен прийняти певне рішення. Чи слід лікувати пацієнта? Якщо так, то за допомогою яких ліків чи процедур? Амбулаторно чи стаціонарно? Можна заперечувати відповідність останніх запитань темі валідності вимірювання. Але, з практичної точки зору, аспект прийняття рішень на основі результатів вимірювання вигідно включити до загального дослідження валідності вимірювання, інакше нам доведеться розглядати окремо «валідність прийняття рішень». А тепер задамося останнім запитанням: якщо ми забезпечили перевірку усіх перелічених свідчень валідності вимірювання температури тіла, чи можемо ми, нарешті. стверджувати, що вимірювання є цілком валідним? На жаль, ствердну відповідь ми можемо дати лише з тією чи іншою мірою впевненості, щоправда, ніколи не повною. Можна наводити скільки завгодно аспектів, які випали з нашого розгляду. Наприклад, чи був збитий ртутний стовпчик до достатньо низького рівня перед початком вимірювання температури?

Розглянутий приклад свідчить про необхідність комплексної валідизації навіть такого простого, у порівнянні з психометричним, фізичного вимірювання. Але для більшості фізичних вимірювань згадані вище проблеми й методи їх вирішення є цілком очевидними, тому й термін «валідність» для них зазвичай не використовується.

Психометричні вимірювання, до яких ми відносимо й освітні, відрізняються від фізичних кардинально. Головними відмінностями психометричних вимірювань є латентність вимірюваних величин, і те, що отримані результати є практично у кожному конкретному випадку частинними й потребують ряду узагальнень.

Розробник чи користувач тесту ніколи не може із стовідсотковою упевненістю сказати, що саме вимірюється даним тестом. Якщо ми хочемо визначити вагу власного тіла, у нас не виникає сумнівів, що для цього слід скористатися вагами, а не, скажімо, лінійкою. З іншого боку, коли у нас в руках опиняється інструмент на зразок ваг, ми можемо впевнено стверджувати, що може і чого не може вимірювати цей інструмент. З тестами все виглядає куди складніше. Наприклад, тест на *рівень інтелекту*, за умови значного скорочення часу на його виконання, може перетворитися на тест *здатності до концентрації розумових зусиль*, і користувач тесту може такого перетворення не помітити. У тесті досягнень з історії можуть зустрічатися настільки великі уривки тексту, що тест вимірюватиме радше швидкість читання, ніж власне знання історії.

Джерел поганої валідності психологічних вимірювань є багато, вони пов’язані з самою природою тестування. Основною причиною можливого погіршення валідності є вибірковий метод – звуження цільової популяції (людей, проявів їх поведінки) до множини її окремих представників, з наступним узагальненням отриманих результатів на всю популяцію.

В стандартній процедурі тестування подібне звуження відбувається тричі (рис. 4.1):

1. З області поведінки, яка досліджується (target domain) обираються для тестування лише деякі прояви – ті, які в принципі можуть бути перевірені тестом. Вони складають популяцію, або генеральну сукупність, тестових завдань (не слід її плутати з поняттям банку завдань).

2. Для тесту з популяції тестових завдань виконується вибірка завдань.

3. Тест отримує свої характеристики після апробації лише на вибірці з цільової популяції осіб, тобто тих осіб, для яких тест призначений.

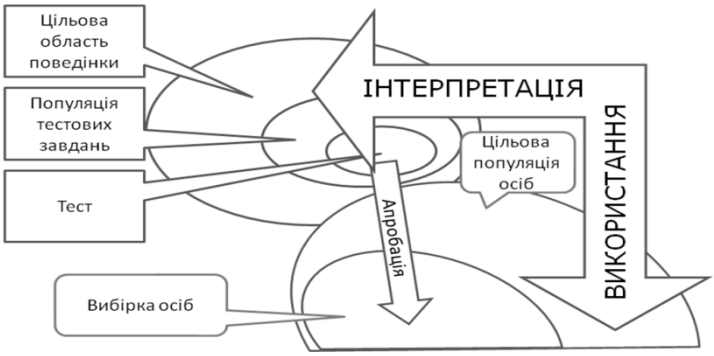


Рис. 4.1. Співвідношення множин суб'єктів та об'єктів тестування

Але цим не обмежується коло джерел можливої «поганої» валідності тесту. Результати одного і того ж тесту, отримані на одній і тій же вибірці з цільової популяції суб'єктів, можна порізному *інтерпретувати*. Неправильна інтерпретація результатів тестування перекреслює всі попередні зусилля, направлені на забезпечення валідності тесту. З іншого боку, теоретично можна припустити, що для деякого новоствореного тесту може існувати така, хай і невідома, інтерпретація його результатів, яка зробить тестування валідним, і тоді постає завдання відшукати цю правильну інтерпретацію.

Навіть після правильної інтерпретації результатів тестування можуть бути прийняті неправильні *рішення*. Наприклад, якщо дитина не пройшла тест на готовність відвідувати дитячий садок, і було прийнято рішення відкласти прийом до садка до наступного року, це рішення може бути як правильним, так і неправильним, в залежності від того, з яких саме причин дитина не склала тест успішно. Не всі дослідники погоджуються з тим, що подібні проблеми якось стосуються валідності тестування. Але, включаючи про-

блеми правильності прийнятих на основі тестування рішень до кола проблем валідизації, ми вчинимо розумно, бо тим самим ми не залишимо цих проблем поза увагою і зменшимо ризик прийняття неправильних рішень.

Щойно сказане стосується також і можливих *соціальних наслідків* тестування. Тестування, особливо широкомасштабне, доленосне для його учасників, часто супроводжується збуреннями в суспільстві, породжуючи напругу, скандали, міфи, непорозуміння, політичні спекуляції. Було б правильним, якби розробники і користувачі тестів зважали на можливі суспільні ефекти і включали дослідження соціальних наслідків тестування до кола питань валідизації вимірювань.

Говорити «валідність тесту» некоректно у тому розумінні, що потрібно уточнювати, як будуть проінтерпретовані і використані результати тестування, які наслідки вони будуть мати. Тому більш правильно говорити про *валідність як про відповідність передбачуваної інтерпретації та використання результатів тестування тій меті, заради якої створено тест*.

Традиційно виділяють кілька видів валідності. З 20-х по 50 роки минулого століття в теорії валідизації панівну роль відігравало поняття критеріальної валідності. *Критеріальна валідність* визначає ступінь відповідності результатів тестування зовнішнім, тобто таким, які не стосуються тесту, критеріям. Наприклад, критерієм для тесту особи на здатність до навчання в університеті може бути середній бал за 1 курс, отримані цією особою після вступу в університет. Зрозуміло, що сам критерій повинен бути валідним. Критеріальна валідність визначається чисельно як коефіцієнт кореляції між тестом і критерієм, обчисленим для репрезентативної вибірки осіб, і називається у цьому випадку *коефіцієнтом валідності*. З точки зору відмінностей за часовою ознакою, критеріальну валідність поділяють на поточну (конкурентну) та прогностичну. *Поточна* валідність отримується при порівнянні результатів тестування з уже відомими на момент тестування результатами. Якщо критерієм є інший тест, то слід обґрунтувати використання нового тесту. Наприклад, слід показати, що новий тест у порівнянні з критерієм є коротшим, зручнішим, чи має якісь інші переваги. Якщо кореляція між новим тестом і критерієм є дуже високою, це може ставити під сумнів необхідність викорис-

тання нового тесту, адже він не даватиме нової у порівнянні з критерієм інформації. Якщо ж кореляція між тестом і валідним критерієм є надто низькою, то тест не може бути визнаним валідним.

Прогностичну валідність розглядають у випадках, коли тест призначений для передбачення рівня успішності особи в певному виді діяльності в майбутньому. Такими є тести здібностей та тести відбору. У цих випадках коефіцієнт валідності обчислюється як кореляція між результатами тестування групи осіб і критерієм, який виражається в оцінці реальної діяльності осіб (див. приклад про порівняння результатів вступного тесту з оцінками за 1 курс). Дослідження прогностичної валідності вимагає багато часу, адже після тестування групи осіб потрібно дочекатися того моменту, коли ця група осіб достатньо проявить себе в обраному виді діяльності настільки, щоб за результатами цієї діяльності отримати валідні критеріальні оцінки. Якщо обставини не дозволяють досліджувати валідність тесту таким чином, то можна запропонувати тест групі осіб, яка вже задіяна у даному виді діяльності і для якої вже існують критеріальні оцінки, після чого обчислити кореляцію між результатами тестування і критерієм.

Валідність тесту досягнень прийнято досліджувати шляхом порівняння його змісту із змістом тієї області, для оцінки якої він призначений. У цьому випадку говорять про *змістову валідність* тесту. На відміну від критеріальної валідності, змістова валідність визначається не чисельно, а у вигляді суджень експертів. Змістова валідність повинна забезпечуватися вже на початкових етапах створення тесту шляхом ретельного аналізу цільової області і складання специфікації тесту та описання окремих тестових завдань у такий спосіб, щоб зміст цільової області був представлений у тесті достатньо повно і пропорційно. Ця робота тільки на перший погляд може здатися простою, адже в поняття успішності засвоєння цільової області, наприклад, певної навчальної дисципліни, повинні включатися і необхідні рівні когнітивних процесів. Наприклад, тест, у якому переважають завдання на знання фактів, не дасть змоги повно виявити рівень розуміння цих фактів та їх взаємозв'язків, здатність особи до самостійного критичного аналізу, оцінювання та застосування набутих знань, а ці якості зазвичай входять до переліку педагогічних цілей, які ставляться при викла-

данні навчальних дисциплін, отже, складають зміст цільової області.

Розробка тестів особистості привела до появи поняття *конструктної валідності*. Для подібних тестів часто не існує прийнятних критеріїв, і неможливо однозначно визначити зміст цільової області поведінки.

Конструктна валідність виникла з потреби вимірювати *теоретичні конструкти*. Класичний приклад конструкту – здібність до чогось.

Поняття конструктної валідності ввели Кронбах і Міл. Вони розглядали цей вид валідності як альтернативу критеріальній та змістовій валідності. Конструктна валідність, за Кронбахом і Мілом, повинна була застосовуватися, коли «тест інтерпретується як міра деякого атрибуту чи якості, які не визначені операціонально». Тобто коли не існує ні прийнятного зовнішнього критерію, ні змістового описання даного атрибуту чи якості.

Щоправда, додають Кронбах і Міл, практично для кожного тесту існує потреба визначення, які психологічні конструкти у ньому задіяні.

Спочатку конструктна валідність розглядалася як окремий випадок валідності. Зазвичай розглядалося чотири види валідності як такі, що є пов'язаними з чотирма видами інтерпретації:

- прогностична і конкурентна (різновиди критеріальної валідності);
- змістова валідність;
- конструктна валідність.

Ці види валідності становили разом так звану «троїсту модель валідності». Однак в кінці 1970-х намітилося 2 тренди в подальшому розвитку теорії:

1. Стійкий інтерес до чіткого визначення, які саме види свідчень потрібні для тих чи інших інтерпретацій та використання тестів.
2. Визнання необхідності створення єдиної концепції валідності.

Базою для єдиної концепції валідності стала конструктна валідність. В кінці 1980-х років Мессік розробив розширену модель конструктної валідності як основу (framework) єдиного поняття валідності. Мессік визначає валідність як інтегральне оціночне

судження про ступінь підтримки емпіричними свідченнями і теоретичними міркуваннями адекватності прийнятності висновків і дій, заснованих на тестових балах чи інших видах оцінок.

В [10] М. Кейн виділив три аспекти, у яких конструктна модель вийшла за межі теоретико-залежного контексту, в якому вона була спочатку запропонована:

1. Між 1955 і 1989 роками основний наголос змістився від валідизації тестів до розробки і валідизації пропозицій щодо інтерпретації та використання тестових балів.
2. Конструктна модель валідності потребує більшою мірою теоретичного дослідження, ніж просто емпіричних свідчень.
3. Фокусування конструктної валідності на теорії веде до можливості і потреби оспорювати запропоновану інтерпретацію і розробляти альтернативні інтерпретації.

Валідність як аргумент: підхід М. Кейна. У викладі Кейна, *аргумент валідності* (validity argument) є інструментом загальної оцінки інтерпретації та використання тестових балів, що передбачаються для даного тесту. Головною метою при цьому є відшукання ясних та взаємно узгоджених свідчень «за» або «проти» запропонованих інтерпретацій чи застосувань, і, якщо можливо, свідчень для альтернативних інтерпретацій/застосувань. З цією метою спочатку розробляється *інтерпретативний аргумент* (interpretative argument), який слугує своєрідною канвою для вироблення аргументу валідності. Інтерпретативний аргумент складається з ряду припущень та декларативних висновків.

Підхід Кейна до валідизації можна сформулювати як послідовність кроків, що може ітеративно повторюватися:

1. Пропонується інтерпретація тестових балів в термінах інтерпретативного аргументу.

2. Створюється попередня версія аргументу валідності шляхом відшукання усіх доступних свідчень правдоподібності інтерпретативного аргументу.

3. Детально оцінюється справедливість припущень та висновків.

4. Якщо потрібно, переформульовуються інтерпретативний аргумент та аргумент валідності, після чого повторюється крок 3. Так відбувається доти, доки не свідчення правдивості задекларо-

ваних висновків не стануть достатніми для їх визнання або відхилення.

Цей процес нагадує процес створення теорій у природничих науках. Вже з розглянутого вище прикладу фізичного вимірювання можна зробити такі важливі висновки: 1) дослідження валідності вимірювання має бути комплексним; 2) навіть комплексне дослідження не дає повної гарантії валідності. Подібну ситуацію математик може характеризувати як таку, у якій є лише багато необхідних умов для доведення деякого твердження, і жодної – достатньої. Для фахівців гуманітарної сфери подібна ситуація є цілком звичною. Таким чином, можна розглядати комплексну валідизацію як міні-теорію, засновану на системі практичної аргументації. Розглянемо загальні принципи побудови такої теорії. За С. Тулміним, аргументація – це переважно процес верифікації вже існуючих ідей. Існує шість взаємопов'язаних компонентів процесу аргументації:

1. Твердження (claim). Твердження повинне бути завершеним. Наприклад, якщо хтось намагається переконати, що він є громадянином Великобританії, то його твердженням буде «я громадянин Великобританії»

2. Свідчення (evidence). Це факт, на який посилаються, як на підставу для твердження. Наприклад, особа з попереднього прикладу може підтримати своє висловлювання іншими даними «я народився на Бермудських островах».

3. Підстава (warrant). Висловлювання, що дозволяє перейти від свідчення до твердження. Для того щоб перейти від свідчення «я народився на Бермудських островах» до твердження «я громадянин Великобританії», особа повинна використовувати підстави для усунення розриву між твердженням і свідченням, вказавши, що «людина, народжена на Бермудських островах, юридично може бути громадянином Великобританії».

4. Підтримка (backing). Доповнення, спрямоване на підтвердження висловлювання, вираженого в підставі. Підтримка має бути використана, коли підстава сама по собі не є достатньо переконливою для опонента.

5. Спростування/контраргумент (rebuttal). Висловлювання, що вказує на обмеження, які можуть застосовуватися. Прикладом контраргументу є: «Людина, що народилася на Бермудських ост-

ровах, може легально бути громадянином Великобританії, тільки якщо вона не зрадила Великобританії і не є шпигуном іншої країни».

6. Визначник (qualifier). Слова та фрази, що виражають ступінь впевненості автора у його твердженні. Це такі слова і фрази, як «ймовірно», «можливо», «неможливо», «безумовно», «ймовірно» або «завжди». Твердження «Я безумовно є громадянином Великобританії» несе в собі набагато більшу ступінь впевненості, ніж твердження «Я імовірно є громадянином Великобританії».

Перші три елементи розглядаються як основні, тоді як потреба у трьох останніх виникає не завжди.

Крім загальної філософії науки, на теорію Кейна вплинули також різні методології, поширені в психометрії, зокрема, висновки узагальненої теорії тестування (Generalizability Theory), а також методологія оцінювання програм (Program Evaluation).

Кейн виділяє чотири широкі категорії інтерпретативних аргументів:

1. Інтерпретація для рис особистості (traits). Сюди відноситься наприклад, випадок вимірювання навчальних досягнень учня з певного предмету.

2. Інтерпретація, заснована на теорії. Такого виду інтерпретації вимагають, наприклад, результати тесту здібностей.

3. Якісна інтерпретація. Стосується області якісного (на відміну від кількісного) оцінювання і підходить для валідизації, зокрема, оцінювання учнів під час занять у класі.

4. Процедури прийняття рішень. Прикладом може бути інтерпретативний аргумент для Програми відповідальності NCLB (No Child Left Behind Act) в США.

Інтерпретаційний аргумент для вимірювання риси. Розглянемо детальніше процес побудови інтерпретативного аргументу для випадку, коли тест призначений для вимірювання проявів деякої риси. *Риса* (*trait*) – схильність індивідуума поводитися або діяти певним чином у відповідь на деякий стимул або завдання, за певного набору умов. Під це визначення підходить поняття навчальних досягнень, отже, сюди відноситься найпоширеніший у педагогічному тестуванні вид тестування – тестування досягнень.

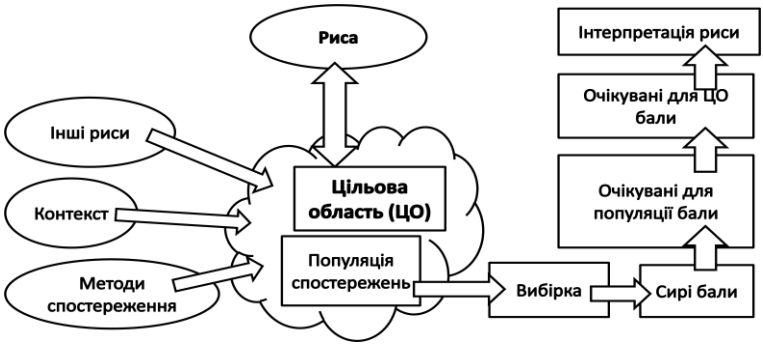


Рис. 4.2. Процедура вимірювання й інтерпретативний аргумент

Риса асоціюється з поняттям *цільової області* можливих спостережень, і очікувані бали особи, які представляють цю особу по відношенню до цільової області, є *цільовими балами*.

Цільова область може бути дуже широкою, як при деяких визначеннях інтелекту, більш вузькою, як при визначенні рівня успішності десятикласника з алгебри, або зовсім вузькою, наприклад, навички з виконання деякого завдання.

На рис. 4.2 схематично зображено взаємозв'язок між процедурою вимірювання і структурою інтерпретативного аргументу у випадку вимірювання риси.

Нехай потрібно дослідити валідність предметного тесту для випадку, коли метою вимірювання є з'ясування рівня успішності (в іншій термінології – рівня навчальних досягнень) особи з даного предмету. Порядок узагальнень інтерпретації тестових балів у цьому випадку є таким (в дужках пропонуються короткі назви відповідних рівнів узагальнення):

1. Від спостережених відповідей – до тестових балів (*скоринг*).

2. Від тестових балів – до балів за ту частину предмету, яку представляв тест (*генералізація*).

3. Від балів з частини предмету, яку представляє тест – до балів з усього предмету (*екстраполяція*).

4. Від балів з предмету – до словесного описання рівня успішності (*імплікація*).

Розглянемо можливі твердження інтерпретаційного аргументу й методи отримання свідчень валідності для кожного з цих рівнів.

Скоринг. Під цим терміном розумітимемо нарахування тестових балів за заданими укладачем правилами. На даному рівні інтерпретаційний аргумент може містити такі твердження:

1. Схема нарахування тестових балів є прийнятною.

2. Схема нарахування тестових балів використовувалась правильно.

3. Нарахування тестових балів було неупередженим.

4. Нараховані бали узгоджується з обраною моделлю шкалування.

Аргумент валідизації для цих тверджень повинен будуватися вже на початку створення тесту. Методи забезпечення валідності для цього етапу мають якісний характер і, здебільшого, форму експертних висновків. Найкраще було б, якби дві або й більше груп експертів виробляли незалежно одна від одної власні схеми оцінювання. Ці різні схеми слід проаналізувати, порівнюючи бали, отримані для тесту за цими схемами. Важливо також правильно організувати та контролювати роботу оцінювачів, якщо така робота диктується видами тестових завдань (наприклад, тест містить есе).

Практично завжди до отриманих «сирих» тестових балів застосовується той чи інший метод шкалування. Наприклад, в Україні тести ЗНО протягом останніх років шкалувалися за методом еквіпроцентильної нормалізації. Доцільність використання цього методу є предметом постійних дискусій. Справедливо вказується головний його недолік – неможливість інтерпретувати тестові бали як дійсний рівень успішності (що знає й уміє випускник, і чого він не знає і не вміє), оскільки у випадку застосування цього методу оцінювання є нормо-орієнтованим. Тим не менше, для порівняння результатів тестування з предмету у різних сесіях ЗНО і за різні роки, цей метод є чи не єдино правильним (крім, хіба що, застосування методів теорії IRT), з огляду на непорівнюваність самих тестів. Зауважимо, що метод еквіпроцентилів рекомендований Європейською Комісією для зарахування наявних оцінок студента на новому місці навчання при його переході до іншого університету. З точки зору забезпечення валідності вимірювання, шкалування

не повинне суперечити тій кінцевій меті, заради якої проводиться тестування. Так, при застосуванні методу еквіпроцентильної нормалізації у ЗНО важливим є відповідність його головному принципу – той випускник, який отримав більше «сирих» балів у порівнянні з іншим випускником, також повинен отримати більше балів і після проведення процедури шкалування.

Генералізація. Практично завжди тестуванням може бути охоплена не вся область поведінки суб'єкта, яка є предметом вимірювання. Це стосується й предметних тестів. Наприклад, якщо тест з іноземної мови проводиться у письмовій формі, неможливо перевірити правильність вимови учня. З іншого боку, тестові завдання завжди є лише вибіркою з усіх можливих тестових завдань, які разом складають деяку генеральну сукупність завдань (за термінологією математичної статистики). Саме ця генеральна сукупність тестових завдань й репрезентує ту частину поведінки особи у цільовій області (у нашому випадку – проявів рівня успішності з предмету), яка покривається тестуванням. Не слід плутати генеральну сукупність тестових завдань з банком тестових завдань. Доцільно вважати, що генеральна сукупність містить нескінченно багато завдань. Адже легко погодитися з тим фактом, що тести двох різних незалежних укладачів, будучи вибірками з генеральної сукупності, практично ніколи не містять однакових завдань.

При просуванні інтерпретації тестових балів від конкретного тесту до генеральної сукупності тестових завдань важливо отримати свідчення на користь таких тверджень (інтерпретаційний аргумент):

1. Завдання тесту складають репрезентативну вибірку з генеральної сукупності тестових завдань.

2. Дана вибірка тестових завдань є достатньо великою, щоб контролювати випадкову похибку вимірювання.

Аргумент валідності на цьому рівні будується методами математичної статистики. Зокрема, контроль похибок вибірок здійснюється методами теорії надійності та узагальненої теорії тестування (Generalizability Theory).

Екстраполяція. На рівні екстраполяції інтерпретація тестових балів переноситься на всю область, яка оцінюється (у нашому випадку – це рівень успішності з даного предмету, наприклад, з математики чи з історії України).

Це досить складний і відповідальний етап у дослідженні валідності. Інтерпретаційний аргумент для цього етапу може містити наступні твердження:

1. Тестові бали є релевантною мірою рівня успішності з даного предмету.

2. Систематичні похибки вимірювання не перешкоджають екстраполяції.

Методи збору свідчень валідності на даному етапі можна умовно поділити на аналітичні й емпіричні.

До аналітичних методів належить, зокрема, перевірка широти покриття тестуванням усього різноманіття проявів рівня успішності з предмету. Зауважимо, що істотно збільшити широту покриття можна, замінивши паперове тестування комп'ютерним. Наприклад, при тестуванні з іноземної мови комп'ютерна форма дозволяє відносно легко організувати аудіювання. Перевірка широти покриття тестом успішності засвоєння предмету передбачає, крім перевірки змістової частини предметної області, також і контроль процесів мислення учня. Слід намагатися порівняти ті процеси мислення, які учень використовує під час виконання тестових завдань, з тими процесами мислення, які він демонструє в тій частині володіння предметом, яка не підлягає тестуванню. Важливу роль на цьому етапі відіграють факти і методи когнітивної психології. Зокрема, для виявлення процесів мислення, застосовуваних учнем під час вирішення ним конкретних завдань, використовується метод документування міркувань уголос.

Певну роль відіграє забезпечення так званої очевидної валідності (face validity) тесту, особливо тоді, коли тест (як у випадку ЗНО) є тестом високої відповідальності.

Окремої уваги заслуговує дослідження негативного впливу стандартизації вимірювання. Заходи стандартизації, маючи на меті зменшення випадкової похибки вимірювання, є одночасно джерелом систематичних похибок, що є результатом сильного звуження усього розмаїття умов і способів, у які проявляється рівень успішності учня з предмету, до жорстких, максимально однакових для всіх змісту і процедури тестування.

До емпіричних методів екстраполяції тестових балів слід віднести порівняння результатів тестування з критерієм (*критеріальна валідність*); з результатами інших видів оцінювання (*конвер-*

гентна валідність); з оцінками дивергентних рис (*дискримінантна валідність*). Зауважимо, що для дослідження критеріальної валідності потрібен валідизований критерій, а такий є у розпорядженні дослідника дуже рідко, тому те, що часто декларується як критеріальна валідність, повинне бути віднесене швидше до конвергентної валідності. Аналіз матриці «багато рис – багато методів (Multitrait-Multimethod*)»*, яка складається з коефіцієнтів кореляції дивергентних рис, кожна з яких виміряна кількома різними способами, є потужним методом дослідження дискримінантної валідності.

Часом може стати у нагоді використання висновків *попередніх досліджень валідності*, якщо нова ситуація (нові умови, нова популяція учнів, новий тест) мало відрізняється від попередньої.

Імплікація. Якщо на попередніх рівнях валідизації валідність тесту перевіряється для тестових балів, то на рівні імплікації предметом дослідження є вже словесна інтерпретація цих балів. Тут потрібно перевірити, чи є прийнятною смислова інтерпретація результатів вимірювання риси або теоретичного конструкту, і чи узгоджуються властивості отриманих результатів з висновками, асоційованими з визначенням риси чи конструкту. Вербальна інтерпретація тестових балів повинна забезпечуватися вже на стадії розробки тесту. Пізніше, за наявності емпіричних даних, перевіряється узгодження смислової інтерпретації результатів вимірювання з тими висновками щодо співвідношення з іншими змінними, які закладаються в концепції даної риси чи конструкту.

Інтерпретаційний аргумент для вимірювання теоретичного конструкту. Якщо предметний тест використовується для вимірювання теоретичного конструкту, яким є, наприклад, здатність особи до навчання, між рівнями екстраполяції та імплікації у дослідженні валідності тесту з'являється ще один рівень: *теоретична інтерпретація*. Риса (рівень успішності з дисципліни), яка вимірюється, у цьому випадку відіграє роль одного з багатьох можливих *індикаторів конструкту*. Теорія, яка спочатку існує лише «в голові» дослідника, передбачає певне співвідношення між індикаторами, а також між даним конструктом та іншими конструктами, і результати тестування повинні або підтверджувати цю теорію, або спростовувати її.

Конструктна валідність досліджується насамперед методами кореляційного аналізу та моделювання структурними рівняннями (Structural Equation Modeling). Останній зараз все частіше використовується і претендує на універсальність. Що стосується самої теорії, яка будується для даного конструкту, то вирішальну роль тут відіграють досягнення когнітивної психології, методи моделювання когнітивних процесів.

Підсумовуючи, підкреслимо важливість комплексного дослідження валідності тестів, зокрема, предметних, і, як наслідок, необхідність логічного впорядкування цього дослідження. Оскільки комплексна валідизація вимірювання є різновидом системи практичної аргументації, то підхід, який базується на побудові інтерпретаційного аргументу й відповідного йому аргументу валідності, є запорукою її повноти та високої якості.

5. НАДІЙНІСТЬ

Загальне поняття надійності. Поняття надійності, разом з поняттям валідності, є фундаментальною характеристикою тесту, без якої тестування не може вважатися вимірюванням. Разом з тим, у порівнянні з валідністю, надійність є більш технічною характеристикою, яка стосується насамперед проблеми точності вимірювання. У повсякденній мові ми називаємо надійним помічником або надійним другом людину, на яку можна покластися у певній складній ситуації, тобто людину, дії якої у визначених умовах є цілком прогнозованими. Подібно до цього, тест вважається надійним, якщо його багаторазове використання у схожих умовах приводить до схожих же результатів. Ця цілком зрозуміла з практичної точки зору вимога, однак, не може вважатися строгим визначенням надійності, оскільки поняття схожості можна надто вільно трактувати: схожість може бути більшою або меншою, сильнішою або слабшою. Проте ми не можемо замінити тут слово «схожі» на «однакові», оскільки після такої заміни практична перевірка виконання цієї вимоги стає неможливою.

Справді, навіть якщо один і той же тест пред'являти одній і тій же групі екзаменованих, ми змушені будемо робити це, як мінімум, у різні моменти часу. Але ж психічні процеси, які невпинно протікають у мозку людини, наявність пам'яті, тренованості, здатності до навчання призводять до того, що результати першого тестування неодмінно впливатимуть на результати другого, і ми отримаємо вже для двох сеансів тестування дві різні, хоча, можливо, й схожі ситуації.

Навіть для простих фізичних вимірювань спостерігається щось подібне. Якщо, приміром, зважити з великою точністю кілька разів дерев'яний брусок, то результати різних зважувань будуть дещо відмінними. Іншими словами, вимірюванню властива певна *похибка*. Важливим завданнями є обчислення та ідентифікація джерел похибки вимірювання. Якщо між зважуваннями бруска будуть достатньо великі проміжки часу, і вологість повітря буде при різних зважуваннях різною, істотним джерелом похибки вимірювання буде здатність дерева змінювати свою вологість, і, як

наслідок, вагу. Інші джерела похибок, такі, як ретельність зчитування показів інструменту зважування, при цьому можуть бути такими, що їх середній результат буде близьким до нуля внаслідок взаємної компенсації похибок з від'ємними і додатними значеннями.

Різні схеми організації повторного психометричного вимірювання неодмінно містять різні істотні джерела похибок, які призводять до того, що результати вимірювання щоразу, взагалі кажучи, будуть дещо відмінними. Так, якщо групі осіб двічі пред'являється один і той же тест, таким джерелом мінливості похибки є мінливість часових інтервалів між першим і другим сеансом тестування; якщо групі осіб пред'являються у різні моменти часу різні форми одного і того ж тесту (їх називають паралельними), то з'являється додаткове джерело похибки – відмінність у змісті завдань паралельних форм; якщо результати одиничного тестування оцінюються групою експертів незалежно один від одного, то істотним джерелом похибки є відмінність у вподобаннях та критеріях оцінювання між експертами. Крім того, оскільки тест складається з окремих завдань, узгодженість у результатах між окремими завданнями та завданнями і тестом в цілому теж є предметом надійності.

Відповідно до того, яка схема повторного використання тесту застосовується, можна говорити про різні види надійності. Надійність при цьому будемо шукати у числовому вираженні. Але теоретично більш правильний підхід полягає у тому, що для даного тесту існує деяка ідеальна величина, яка називається коефіцієнтом надійності, а різні схеми практичного дослідження надійності дають різні *оцінки* цього коефіцієнта. Для уведення поняття коефіцієнта надійності далі розглянемо так звану класичну модель тестової оцінки. Цю модель, разом із деякими припущеннями, та висновками, які з них випливають, називають класичною теорією тестування (*CTT – Classical Test Theory*). Пізніше познайомимося з так званою узагальненою теорією (Generalizability Theory), яка дозволяє в окремих випадках ідентифікувати вплив різних джерел похибок вимірювання.

Класична модель тестової оцінки. Розробка моделі істинної оцінки опитуваного була розпочата ще Чарльзом Спірменом (Charles Spearman), а потім продовжена різними дослідниками.

Центральним положенням класичної теорії тестування є твердження про те, що спостережена тестова оцінка X_{pf}, яку отримав екзаменований p в результаті виконання ним форми f даного тесту, є сумою двох складових – істинної оцінки екзаменованого T_p та похибки вимірювання E_{pf}:

$$X_{pf} = T_p + E_{pf}.$$

Істинна оцінка особи T_p відповідає її рівню вираженості вимірюваної якості і є незмінною для різних форм тесту. Форми тесту тут вважаються *строго паралельними*, тобто вони задовольняють наступним чотирьом вимогам.

1. Вони мають ідентичні специфікації.

2. Розподіли спостережених оцінок при пред'явленні різних форм різним однаковим за об'ємом репрезентативним вибіркам з популяції екзаменованих є однаковими:

$$F(X_f) = F(X_g) = F(X_h) = \cdots$$

3. Результати пред'явлення цим вибіркам екзаменованих будь-яких двох форм мають однакову коваріацію:

$$s_{X_f X_g} = s_{X_f X_h} = s_{X_h X_g} = \cdots$$

4. Якщо Z – деяка міра тієї ж самої або іншої якості осіб, коваріація результатів пред'явлення різних форм з Z є однаковою:

$$s_{X_f Z} = s_{X_g Z} = s_{X_h Z} = \cdots$$

Наступне припущення полягає у тому, що якщо екзаменований складає повторно різні строго паралельні форми тесту за умови, що попереднє тестування жодним чином не впливає на наступне (повне стирання з пам'яті), середнє значення похибок вимірювання, за умови наближення сеансів тестування до безмежності, прямує до нуля:

$$\mathrm{E}_f(\mathrm{E}_{pf}) = 0.$$

(тут E_f означає математичне очікування по множині строго паралельних форм тесту).

Ще одне припущення полягає у тому, що якщо будь-яка строго паралельна форма тесту пред'являється групі екзаменованих, очікуване середнє похибок вимірювання наближається до нуля при прямуванні кількості екзаменованих до нескінченності:

$$\mathrm{E}_p\left(\mathrm{E}_{pf}\right) = 0.$$

(тут E_p означає математичне очікування по множині усіх екзаменованих).

З цих припущень випливає, що коваріація між істинними балами та похибками вимірювання для будь-якої з паралельних форм тесту дорівнює нулю, і коваріація між похибками вимірювання для будь-яких двох паралельних форм тесту теж дорівнює нулю. Іншими словами, між істинними балами і похибкою вимірювання при адмініструванні однієї форми тесту, а також між похибками вимірювання при адмініструванні різних форм тесту існує лінійна незалежність. З іншого боку, істинні оцінки та похибки вимірювань як компоненти спостережених оцінок корелюють з ними.

З того, що істинні оцінки екзаменованих і похибки вимірювання некорельовані, випливає той ключовий факт, що для окремої форми тесту дисперсія спостережених оцінок є сумою дисперсій істинних оцінок і похибок вимірювання:

$$s^2\left(X_f\right) = s^2(T) + s^2\left(E_f\right).$$

Також з припущень класичної моделі виливає, що коваріація між спостереженими оцінками, отриманими за паралельні форми тесту, дорівнює дисперсії істинних оцінок:

$$s_{X_f X_g} = s_T^2.$$

Поділивши цю рівність на добуток середньоквадратичних відхилень спостережених балів за відповідними формами тесту, отримаємо коефіцієнт кореляції:

$$\rho_{X_f X_g} = \frac{s_{X_f X_g}}{s_{X_f} s_{X_g}} = \frac{s_T^2}{s_X^2} = \frac{s_T^2}{s_T^2 + s_E^2}.$$

Цю величину назвемо коефіцієнтом надійності, або просто надійністю. Таким чином, *коефіцієнт надійності* – це коефіцієнт кореляції між оцінками за гіпотетичні строгі паралельні форми тесту. Іншими словами, це відношення дисперсії істинної оцінки до дисперсії спостереженої оцінки.

Ще один важливий факт випливає також з наведених рівностей: коефіцієнт надійності дорівнює квадрату коефіцієнта кореляції між істинними та спостереженими оцінками при однократному тестуванні. Як зазначалося раніше, квадрат коефіцієнта кореляції між двома змінними визначає частку дисперсії, яку одна змінна привносить у дисперсію іншої змінної. Таким чином, величина

$$1 - \rho_{X_f X_g} = 1 - \rho_{XT}^2$$

визначає частку загальної дисперсії спостережених оцінок, зумовлену дисперсією похибки вимірювання.

Стандартна похибка вимірювання. Зауважимо, що припущення класичної теорії тестування не вимагають, щоб для двох різних осіб, яким пред'являлися повторно різні паралельні форми тесту, мінливість спостережених балів була однаковою. Теорія і практика вимірювань свідчать, що мінливість оцінок за різні форми тесту з завданнями, виміряними за дихотомічною шкалою, є меншою для осіб з дуже великими або дуже малими істинними оцінками, ніж для осіб з істинними оцінками, близькими до центру розподілу. Стандартне відхилення s_{T_p} для осіб з рівнем T_p називається *умовною стандартною похибкою вимірювання*. В свою чергу, безумовна *стандартна похибка вимірювання* визначається як корінь квадратний з середньої очікуваної по всій групі екзаменованих дисперсії похибки вимірювання. Це означає, що стандартна похибка вимірювання може бути визначеною лише відносно певного розподілу істинних оцінок. Для вибірок з різних популяцій

екзаменованих ця величина, взагалі кажучи, буде різною. Тому називати її безумовною можна лише з урахуванням цієї обставини.

Між безумовною стандартною похибкою вимірювання, коефіцієнтом надійності та дисперсією спостережених оцінок існують такі співвідношення:

$$s_E = \sqrt{s_X^2\left(1 - \rho_{X_f X_g}\right)};$$

$$\rho_{X_f X_g} = 1 - \frac{s_E^2}{s_X^2};$$

$$s_X^2 = \frac{s_E^2}{1 - \rho_{X_f X_g}}.$$

Таким чином, при наявній дисперсії спостережених оцінок, надійність тесту можна трактувати як через поняття коефіцієнта надійності, так і за допомогою поняття стандартної похибки вимірювання.

На відміну від коефіцієнта надійності, стандартна похибка вимірювання дає ту перевагу, що дозволяє оцінювати точність оцінки учасника тестування. Якщо припустити, що оцінки опитуваного при тестуванні за допомогою паралельних форм тесту будуть розподілені рівномірно, то середнє цих оцінок досить точно відповідатиме істинній оцінці, а стандартна похибка вимірювання буде стандартним відхиленням цього розподілу. Відомо, що при цьому розподіл самої стандартної похибки буде близьким до нормального, і це дає змогу будувати довірчий інтервал з заданою мірою довіри (див. главу 2) для отриманої оцінки. Центром цього інтервалу фактично є, як ми сказали, істинна оцінка у вигляді середньої оцінки по всіх паралельних формах, а радіус інтервалу визначається для різних значень надійності властивостями нормального розподілу. Так, можна з упевненістю близько 68% (довірчою ймовірністю 0,68) стверджувати, що спостережена оцінка опитуваного буде відхилятися від істинного бала не більше, ніж на одиницю стандартної похибки. Припустимо, що істинна оцінка опитуваного дорівнює 50, а стандартна похибка вимірювання дорівнює 2. Тоді з упевненістю 68% можна стверджувати, що спостережений бал опитуваного знаходиться в межах від 48 до 52

балів, або з упевненістю 95%, що він знаходиться у межах від 46 до 54 балів (тобто не далі ніж на дві одиниці стандартної похибки від істинної оцінки).

Але проблема практичного застосування цього факту полягає у тому, що істинна оцінка опитуваного нам не відома, а відома, навпаки, спостережена оцінка. Чи можемо ми так само побудувати довірчий інтервал для істинної оцінки з центром у значенні спостереженої оцінки? Строго кажучи, ні. Тим не менше, можна скористатися формулою, запропонованою Галліксеном у 1950 році, яка дозволяє визначати для окремого опитаного інтервал з центром у істинній оцінці з упевненістю 68%:

$$I = \bar{X} + \rho_{XX'}(X - \bar{X}) \pm s_X\sqrt{1 - \rho_{XX'}}\sqrt{\rho_{XX'}},$$

де $\bar{X}$ – середнє арифметичне оцінок групи опитаних, $\rho_{XX'}$ – коефіцієнт надійності (точніше коефіцієнт альфа Кронбаха, див. далі у цій главі), X – спостережена оцінка опитаного, s_X – стандартне відхилення оцінок у групі опитаних. Тут у лівій частині рівності вираз перед знаком $\pm$ відповідає істинній оцінці опитаного, а після цього знаку – стандартній похибці вимірювання. Нехай, наприклад, Володимир отримав оцінку 79 за тест з математики, а в документації до тесту сказано, що для групи у 1200 осіб з цільової популяції, до якої належить і сам Володимир, середня оцінка становила 73, стандартне відхилення – 9, а коефіцієнт надійності альфа – 0,93.

Тоді за наведеною формулою отримаємо $I = 78{,}6 \pm 2{,}3$. Отже, можна стверджувати з упевненістю 68%, що істинна оцінка Володимира знаходиться в інтервалі між 76,3 і 80,9 і її можна вважати рівною 78,6. Порівнявши знайдену істинну оцінку із спостереженою (79), бачимо, що вона дещо нижча. Причиною цього є той факт, що спостережена оцінка Володимира вища від середньої по репрезентативній вибірці. Якби спостережена оцінка була нижчою від середньої, то істинна оцінка, навпаки, була би дещо вищою. Використовуючи таблицю значень для нормального розподілу, ми також можемо відповідати на питання на зразок наступного: наскільки ймовірно, що істинна оцінка становить 80 або більше? Важливість подібного питання є очевидною, якщо, скажімо, величина у 80 балів є пороговою для прийняття рішення, напри-

клад, прийому Володимира до класу з поглибленим вивченням математики.

Ще одна проблема полягає у тому, що значення стандартної похибки, отримане на основі даних про опитування усієї групи, не однаково добре підходить у різних точках розподілу оцінок членів групи. Було показано, що стандартна похибка вимірювання є найбільшою в середній області шкали оцінок, і значно меншою – на кінцях цієї шкали. Різниця може досягати двох і більше разів. Для гомогенної групи осіб (тобто осіб з однаковими істинними оцінками) дисперсія похибки дорівнює

$$\sum_j P_j(1 - P_j),$$

де P_j – рівень трудності j-го завдання (тобто ймовірність правильної відповіді на це завдання для представника цієї групи, якщо, завдання оцінюється за дихотомічною шкалою «правильно-неправильно»).

Існує простий спосіб, знайдений емпірично, який дозволяє оцінити стандартну похибку вимірювання через кількість завдань у тесті ще до початку його застосування: потрібно знайти корінь квадратний з кількості завдань у тесті і помножити його на 0,45, якщо тест має середню складність, або помножити на 0,3, якщо тест легкий, із середнім балом близько 90%.

Відмінність між поняттями надійності та валідності. Як ми вже зазначали, надійність тесту стосується точності вимірювання. Натомість, валідність пов'язана з самою природою атрибутів, які вимірюються. Не валідний тест може виявитися цілком надійним. Припустимо, що групі екзаменованих з математики помилково пред'являвся тест з мови. І внутрішня узгодженість тестових завдань, і повторне пред'явлення цього тесту можуть вказувати на високу надійність, тоді як змістова валідність його, очевидно, є абсолютно неприйнятною. З іншого боку, не надійний тест ніколи не може вважатися валідним, тому в схему комплексного дослідження валідності тесту входить також і дослідження його надійності.

Надійність тесту стосується лише ситуацій, у яких інструменти вимірювання з подібною якістю застосовуються у подібний спосіб. Валідність же має справу також із співставленням результатів даного вимірювання з результатами альтернативних вимірювань тієї самої риси чи конструкту, або й навіть з результатами вимірювання інших якостей (дискримінантна валідність). Кореляція між двома альтернативними тестами з мови може розглядатися як оцінка надійності тесту – у тій мірі, у якій альтернативні тести можуть вважатися паралельними формами одного тесту. Повторне тестування за допомогою одного і того ж тесту однієї і тієї ж групи осіб дасть іншу оцінку коефіцієнта надійності. Але кореляція між оцінками за два цілком різні тести, скажімо, розробленими різними тестовими компаніями, вже не може розглядатися як оцінка коефіцієнта надійності, тобто різні тести, призначені для вимірювання нехай навіть і одного й того ж конструкту, не можуть розглядатися як повторне використання одного інструменту вимірювання.

Разом з тим, надійність, як і валідність, стосується не стільки самого тесту, скільки очікуваної інтерпретації та використання його результатів.

Розглянемо тепер різні схеми дослідження надійності тесту, які дозволяють знайти різні оцінки коефіцієнта надійності.

Випадок повторного тестування: ретестова надійність. Найбільш очевидний метод оцінювання надійності вимірювання – його повторне застосування до однієї і тієї ж групи осіб. У цьому випадку обчислюється коефіцієнт кореляції між двома отриманими вибірками результатів. Оскільки обидва рази використовується той самий тест, не виникає проблем із забезпеченням строгої паралельності форм тесту.

Але виконати два вимірювання для одних і тих же суб'єктів одночасно неможливо. Тому потрібно визначити, яким має бути проміжок часу між двома тестуваннями. Саме тривалість цього проміжку є тим додатковим джерелом похибки вимірювання, який впливає на точність вимірювання і на показники надійності тесту. В свою чергу, на вибір проміжку часу між тестуваннями сильно впливає те, яка саме якість вимірюється. Деякі психологічні конструкти є цілком стійкі у часовому вимірі, тому для них вплив часового інтервалу між тестуваннями є неістотним. Для інших якос-

тей характерна помірна мінливість у часі. Наприклад, коефіцієнт інтелекту (IQ) для дітей мало змінюється протягом такого періоду, як дошкільний, і два тестування IQ протягом цього періоду можуть показати високу надійність вимірювання. Якщо ж одне тестування провести у дошкільному віці, а інше – у старшому шкільному, або у дорослому віці, слід очікувати набагато більш слабкої кореляції результатів. Взагалі кажучи, при повторному вимірюванні фактично будь-якої якості кореляція результатів із зростанням інтервалу часу між вимірюваннями зменшується.

Нарешті, ретестовий метод зовсім не підходить для оцінки надійності вимірювання навчальних досягнень. Припустимо, що повторно за допомогою одного і того ж тесту вимірюється рівень навчальних досягнень з математики. Під час першого виконання тесту учні мимоволі запам'ятовують відповіді на завдання, у правильності розв'язання яких вони упевнені. При повторному виконанні тесту вони швидко справляються з цими завданнями і мають більше часу для роботи над нерозв'язаними раніше завданнями. Крім того, під час перерви між тестуваннями їх мозок мимоволі, свідомо чи підсвідомо, працював над нерозв'язаними завданнями, і це також збільшує шанси справитися з цими завданнями. Збільшення ж часу між двома тестуваннями, хоча й послаблює ефект запам'ятовування, породжує нові проблеми, адже у цьому випадку учні або продовжували вивчати математику і розвивати свій рівень логіко-математичного мислення, або, навпаки, забували пройдений матеріал і втрачали набуті знання й уміння. Перше призведе до завищеної оцінки істинного значення коефіцієнта надійності вимірювання, друге – до заниженої. В обох випадках істинний рівень навчальних досягнень учнів змінювався.

Оскільки інтервал часу між двома вимірюваннями обирається користувачами тесту довільно, можна формально говорити про вплив вибірки часових інтервалів як джерела похибки при оцінюванні коефіцієнта надійності.

В усіх випадках зменшення часу між двома тестуваннями до нуля, тобто повторне тестування без перерви, породжує принаймні ще одну проблему – вплив втомлюваності опитуваних.

Використання паралельних форм тесту. Нехай тепер у двох сеансах тестування екзаменованим пред'являлися різні (але паралельні) форми тесту. Для тестів навчальних досягнень вплив

вибірки часових інтервалів, особливо коротких, можна вважати у цьому випадку більш слабким. Але цей вплив усе ж залишається, оскільки учні запам'ятовують не тільки конкретні розв'язки завдань з попереднього тестування, але й загальні методи та підходи до їх розв'язання, а вони мають бути спільними для обох форм тесту. Таким чином, паралельні форми тесту в аспекті принципів виконання завдань можуть виявитися однією і тією ж формою. Хоча використання паралельних форм тесту поширене на практиці, оскільки нівелює такі небажані дії, як пряме запам'ятовування відповідей чи списування, слід пам'ятати, що ефект «натаскування» учнів на одній з форм істотно впливає на успішність виконання ними іншої форми. Щоправда, якщо тренованість усіх опитуваних між двома сеансами тестування зросте більш-менш рівномірно, це не вплине на величину коефіцієнта надійності.

При повторному тестуванні за допомогою паралельних форм тесту з'являється принципово нове, у порівнянні з ретестовим методом, джерело похибки оцінки коефіцієнта надійності – *вибірка змісту завдань*. Слід розуміти, що поняття строго паралельних форм тесту, сформульоване вище, є ідеальним, на практиці мало досяжним. Незважаючи на вимогу однакових специфікацій для відповідних завдань у двох паралельних формах тесту, їх зміст все ж є відмінним. Наприклад, один і той же учень може знати, скільки буде 6 помножити на 8, і не знати, скільки буде 6 помножити на 7, хоча формально можна вважати ці два завдання паралельними.

Таким чином, при визначенні оцінки коефіцієнта надійності методом повторного використання паралельних форм тесту існують два істотні джерела похибки – вибірка часових інтервалів і вибірка змісту завдань.

Випадок одноразового тестування. Дослідження надійності вимірювання методами повторного тестування є затратним з точки зору людських та матеріальних ресурсів, а у деяких випадках – взагалі неприйнятним. Оскільки тест складається з багатьох окремих завдань, з'являється ідея штучного поділу тесту на частини, які вважалися б більшою чи меншою мірою паралельними формами тесту. У цьому випадку для дослідження надійності достатньо провести однократне тестування вибірки з цільової популяції осіб. Оцінку коефіцієнта надійності, отриману в такий спосіб, називають оцін*кою внутрішньої узгодженості тесту*.

В усіх випадках оцінювання коефіцієнта надійності за допомогою результатів одноразового тестування тест ділиться на частини і обчислюється кореляція між результатами виконання цих частин репрезентативною вибіркою осіб.

Оскільки обрані частини тесту не можуть бути строго паралельними формами, вводиться поняття *еквівалентних* частин. Це поняття означає послаблення вимог порівняно з поняттям строгої паралельності. Пригадаємо, що поняття строгої паралельності форм тесту включає чотири вимоги: однаковість специфікацій завдань, ідентичність розподілів спостережених оцінок, однаковість коваріації між усіма парами форм, однаковість коваріації між кожною з форм та результатами іншого вимірювання. Були запропоновані різні означення еквівалентності форм, кожне з яких тією чи іншою мірою послаблює ці вимоги.

Першим таким означенням було уведене Лордом та Новіком (Lord & Novick, 1968) поняття *τ-еквівалентності*. Це поняття залишає у силі вимогу незмінності істинної оцінки екзаменованого по множині всіх форм, але не вимагає, щоб вимірювання за дома формами виконувалося з однаковою точністю – дисперсія похибки може бути для різних форм різною. Прикладом є дві форми, які відрізняються лише кількістю завдань. Для таких форм виконуються лише дві останні з чотирьох вимог до строго паралельних форм. Розподіли спостережених оцінок за цими формами мають однакові очікувані середні значення, але можуть мати різні очікувані дисперсії.

Близьким до поняття *τ-еквівалентності* є запропоноване тими ж авторами поняття *істотної τ-еквівалентності*. Це поняття допускає, що для всіх екзаменованих існує константа, на яку відрізняються їх істинні бали, отримані за різними двома формами. Оскільки значення коваріації не чутливе до зміни значень однієї з вибірок на одну й ту ж константу, то й у цьому випадку залишаються виконаними останні дві з чотирьох вимог до строго паралельних форм. При цьому розподіли спостережених оцінок за цими формами можуть мати різні очікувані середні значення і різні очікувані дисперсії.

Однорідні форми (Jöreskog, 1971) – ще одне поняття, яке означає послаблення вимог паралельності. У доповнення до припущень істотної *τ*-еквівалентності, це поняття допускає також іс-

нування константи-множника, на який відрізняються істинні оцінки, отримані за двома формами:

$$T_{pf} = b_{fg}T_{pg} + C_{fg},$$

де T_{pf} – істинна оцінка особи p за виконання форми f, T_{pg} – істинна оцінка особи p за виконання форми g, b_{fg} та C_{fg} – відповідні константи. Іншими словами, для однорідних форм тесту допускається лінійна залежність між істинними балами екзаменованих.

Зауважимо, що однорідність означає також істотну τ-еквівалентність, яка, в свою чергу, включає у себе звичайну τ-еквівалентність.

Формули Спірмена-Брауна та Рюлона-Гуттмана. У 1910 році Спірмен та Браун запропонували формулу для оцінки коефіцієнта надійності за двома строго паралельними половинами тесту X_1 і X_2:

$$\rho_{XX'} = \frac{2\rho_{X_1X_2}}{1 + \rho_{X_1X_2}}.$$

Тут і далі $\rho_{XX'}$ означає відповідну оцінку коефіцієнта надійності. Формула Спірмена-Брауна враховує той факт, що кореляція оцінюється для тесту, удвічі довшого від його половин.

Альтернативна формула Рюлона-Гуттмана, вже для істотно τ-еквівалентних форм, має вигляд:

$$\rho_{XX'} = 1 - \frac{s^2_{X_1-X_2}}{s^2_X}.$$

Тут в чисельнику дробу стоїть дисперсія різниць між спостереженими балами за відповідні половини тесту, у знаменнику – дисперсія балів за весь тест. Для строго паралельних форм формули Спірмена-Брауна та Рюлона-Гуттмана дають однаковий результат. У інших випадках формула Спірмена-Брауна дає дещо більше значення, ніж формула Рюлона-Гуттмана. Наприклад, якщо дисперсії частин дорівнюють 6 і 8 відповідно, а кореляція між ними 0,7, то

за формулою Рюлона-Гуттмана отримаємо оцінку 0,819, а за формулою Спірмена-Брауна – 0,824. Припущення про істотну τ-еквівалентність зазвичай є більш прийнятним, ніж про строгу паралельність, але у цьому випадку використання формули Спірмена-Брауна не є достатньо обґрунтованим. Для однорідних форм не існує строгої формули оцінки коефіцієнта надійності. Такі формули існують лише при додаткових припущеннях. Так, припустивши, що дисперсії істинних оцінок і похибки вимірювання є такими, які б одержувалися лише внаслідок простої зміни довжин частин тесту, для цих части можна обчислити ефективні довжини:

$$\lambda_1 = \frac{s_{X_1}^2 + s_{X_1X_2}}{s_X^2}, \quad \lambda_2 = 1 - \lambda_1.$$

Тоді справедлива формула Ангофа-Фелдта:

$$\rho_{XX'} = \frac{4s_{X_1X_2}}{s_X^2 - \frac{(s_{X_1}^2 - s_{X_2}^2)^2}{s_X^2}}.$$

Рекомендується для випадку, коли відношення дисперсій спостережених балів за частини тесту (більшої до меншої) не перевищує 1,15, використовувати просту в обчисленні формулу Спірмена-Брауна, хоча формула Рюлона-Гуттмана є більш прийнятною, а для випадку, коли це відношення знаходиться у межах між 1,15 та 1,30, використовувати формулу Рюлона-Гуттмана. Якщо ж це відношення є більшим від 1,30, слід використовувати формулу Ангофа-Фелдта.

Методи поділу тесту на дві частини. При визначенні, які завдання до якої частини тесту слід віднести для дослідження внутрішньої узгодженості, потрібно керуватися двома основними принципами. По-перше, слід добиватися максимальної паралельності частин. Зазвичай тест складається з завдань, розташованих у порядку зростання їх трудності. У цьому випадку буває достатнім простий поділ тесту за принципом: завдання з парними номерами

відносяться до однієї частини, завдання з непарними номерами – до іншої частини.

Галіксен (Gulliksen, 1950) описує наступний метод поділу тесту. Завдання зображуються точками на площині відповідно до їх трудності та коефіцієнта кореляції між завданням та тестом в цілому. Завдання, близькі між собою візуально, групуються у пари чи більші кластери. Далі всередині кожного кластеру завдання випадковим чином розподіляються до частин тесту. Цю процедуру за потреби можна виконувати окремо для певних частин цільової області вимірювання або різних форматів тестових завдань, щоб максимально забезпечити паралельність як у аспекті контенту, так і у статистичних властивостях.

По-друге, якщо у тесті є групи завдань, об'єднані спільним змістом, кожну групу слід всю відносити до однієї з частин. Наприклад, якщо тест на ефективність читання складається з кількох текстів, після яких ідуть групи завдань, що стосуються цих текстів, то віднесення завдань однієї групи до різних частин тесту може призвести до штучного завищення кореляції, якщо припустити, що різні групи завдань відносяться до різних частин цільової області вимірювання. Якщо всі завдання груп віднести до однієї, тієї чи іншої частини тесту, це може збільшити дисперсію частин, але не коваріацію між частинами, що дозволяє інтерпретувати вплив вибірки змісту як джерело похибки вимірювання.

Поділ тесту більше ніж на дві частини. Формули К'юдера-Річардсона та альфа Кронбаха. У більшості випадків для дослідження внутрішньої узгодженості тесту краще ділити тест більше ніж на дві частини. Якщо у тесті немає зв'язаних однаковим змістом груп завдань, то тест бажано ділити на максимальну кількість частин – по одному завданню у частині.

Потрібно враховувати, що чим більш однорідною є цільова область вимірювання, тим більшою внутрішньою узгодженістю повинен володіти тест. Наприклад, для тесту, який перевіряє лише уміння учнів множити числа, внутрішня узгодженість має бути більшою, ніж для тесту на всі арифметичні операції.

Для тесту з дихотомічними відповідями на завдання К'юдеру та Річардсону належить кілька формул оцінки коефіцієнта внутрішньої узгодженості, з яких найбільш часто використовується так звана формула *KR-20*:

$$\rho_{XX'} = \frac{n}{n-1} \cdot \frac{s_X^2 - \sum_{i=1}^{n} p_i q_i}{s_X^2},$$

де n – кількість завдань у тесті, s_X^2 – дисперсія оцінок за тест, p_i і q_i – частки тих, хто справився і, відповідно, не правився з i-тим завданням.

Можна математично довести, що оцінка коефіцієнта надійності, отримана за формулою KR-20, дорівнює середньому оцінок, отриманих при поділі тесту на дві частини усіма можливими способами. Оскільки при розщепленні тесту на дві частини обирають такий поділ, який забезпечував би максимальну паралельність частин, то формула KR-20 дає дещо нижчий результат, ніж формули для поділу тесту на дві частини, описані вище. Таким чином, різниця між значеннями, знайденими за цими формулами, та формулою KR-20, є показником неоднорідності тесту.

Формула KR-20 у наведеному нами вигляді не підходить для тестів з політомічними відповідями (тобто такими відповідями, які можуть вважатися частково правильними). Для цього випадку існує більш універсальна формула, яку прийнято називати формулою *альфа Кронбаха*:

$$\rho_{XX'} = \frac{n}{n-1} \cdot \frac{s_X^2 - \sum_{i=1}^{n} s_i^2}{s_X^2},$$

де s_i^2 – дисперсія оцінок, отриманих за i-те завдання тесту.

Саме альфа Кронбаха як показник надійності обчислюється та публікується для тестів зовнішнього незалежного оцінювання в Україні.

Надійність оцінювача. В деяких тестах, таких як тести креативності чи проективні особистісні тести, а також у тестах навчальних досягнень з завданнями з розгорнутою відповіддю, велику роль відіграє суб'єктивізм оцінювача. Надійність оцінювача можна визначити, організувавши оцінювання двома незалежними фахівцями. Між двома наборами оцінок, виставлених цими фахівцями за тест, обчислюється звичайний коефіцієнт кореляції. Джерелом

цієї похибки оцінки коефіцієнта надійності у цьому випадку є вибірковість оцінювачів.

Загальний огляд оцінок коефіцієнта надійності. Різні методи оцінки коефіцієнта надійності можна класифікувати у відповідності до кількості необхідних сеансів тестування та форм тесту. Подамо цю класифікацію у вигляді таблиці:

Таблиця 5.1. Класифікація методів оцінки надійності

Кількість тестувань	**Кількість форм тесту**	
	Одна	**Дві**
Одне	1) Метод розщеплення на еквівалентні половини 2) Формула К'юдера-Річардсона	4) Метод паралельних форм (безпосередній)
Два	3) Ретестовий метод	5) Метод паралельних форм (з часовим інтервалом)

Будь-яку оцінку коефіцієнта надійності можна інтерпретувати у частках дисперсії, спричиненої різними джерелами. Так, величина оцінки 0,85 означає, що 85% дисперсії результатів тестування спричинені мінливістю вимірюваної якості у цільовій популяції осіб, а решта 15% - дисперсією похибок.

Зв'язок різних оцінок коефіцієнта надійності з джерелами похибок наведено у таблиці 5.2.

Покажемо, як спеціально підібраний план дослідження надійності допомагає оцінити вплив різних джерел похибки вимірювання (Анастазі, Урбіна).

Нехай 100 учнів проходили тестування на креативність двічі з інтервалом у два місяці за допомогою паралельних форм тесту. Нехай оцінка коефіцієнта надійності за методом паралельних форм (з часовим інтервалом) складала 0,7. Нехай також метод еквівалентних половин за формулою Спірмена-Брауна дав для обох форм оцінку 0,8; у оцінюванні був задіяний додатковий експерт, який оцінював відібрану навмання половину учнівських робіт, і надійність оцінювача складала 0,92. Тоді вплив часового інтервалу та вибірковості змісту дають 1 – 0,7 = 0,3. Тут 1 означає 100% дисперсії. З іншого боку, вплив лише вибірковості змісту дорівнює

1 – 0,8 = 0,2. Звідси отримуємо що 0,3 – 0,2 = 0,1 – вплив вибірковості часу між тестуваннями. Вплив заміни оцінювача дорівнює 1 – 0,92 = 0,08.

Таблиця 5.2. Зв'язок між методами оцінки надійності та джерелами похибок вимірювання

Вид оцінки коефіцієнта надійності	**Джерела дисперсії похибок**
1) Ретестовий	Часова вибірковість
2) Паралельних форм (безпосередній)	Вибірковість змісту
3) Паралельних форм (з часовим інтервалом)	Часова вибірковість плюс вибірковість змісту
4) Еквівалентних половин тесту	Вибірковість змісту
5) KR-20 та альфа Кронбаха	Вибірковість змісту плюс неоднорідність змісту
6) Оцінювача	Відмінність між оцінювачами

Тоді сумарна оцінка дисперсії похибок дорівнює

$$0{,}2 + 0{,}1 + 0{,}08 = 0{,}38.$$

Звідси істинна дисперсія, зумовлена відмінностями у рівні креативності учнів, дорівнює 1 – 0,38 = 0,62.

Елементи теорії генералізації. Загалом дослідження впливу різних джерел похибки вимірювання є предметом розгляду спеціальної теорії, яку в українському перекладі можна назвати *теорією генералізації* (точніше, теорією узагальнюваності – англ. Generalizability Theory), яка заснована на методах спеціального розділу математичної статистики – дисперсійного аналізу. Достатньо детальний вступ до цієї теорії наведено Крокер та Алгіною в [6]. Тут ми познайомимося лише з основними ідеями цієї теорії.

Вимірювання зазвичай розробляється для застосування у певних фіксованих умовах. Але ці умови є проявом більш широкої множини умов, і дослідника може цікавити, наскільки добре ре-

зультати вимірювання можуть бути узагальненими на цю більш широку множину умов. В теорії генералізації набори умов вимірювання називають *фасетами*. Нехай, наприклад, дослідник вивчає уміння дітей писати твори. У вибраних чотирьох випадках кожен учень пише твори на дві різні теми, і всі твори перевіряються трьома експертами. Тоді дизайн дослідження включає три фасети: випадки, теми творів та експертів. Інший приклад: два контролери оцінюють практичні уміння робітників за важких, середніх та легких умов праці. У цьому прикладі є два фасети: контролерів та умов праці. Завданням дослідника є з'ясувати, наскільки результати вимірювання при фіксованих елементах кожного з фасетів можуть бути узагальненими на всі можливі елементи фасетів. Це дослідження узагальнюваності називають *Г-дослідженням* (*G-study*). Воно проводиться для того, щоб на етапі прийняття рішень дослідник міг правильно спланувати (розробити дизайн) відповідного *Р-дослідження* (*D-study*, від англ. *decision* – рішення).

У Р-дослідженні фасет може трактуватися як фіксований або випадковий. Якщо обирається фіксований фасет, то генералізація проводиться тільки в межах тих умов, які з'являються в Р-дослідженні. Якщо фасет випадковий, то умови, які розглядаються в Р-дослідженні, вважаються вибіркою із деякої більшої множини умов.

У класичній теорії тестування істинна оцінка екзаменованого визначається як середнє арифметичне великої кількості результатів строго паралельних вимірювань. Дисперсія істинної оцінки дорівнює дисперсії середніх за результатами паралельних вимірювань, а надійність визначається як відношення дисперсій спостережених та істинних оцінок. В теорії генералізації оцінка екзаменованого в генеральній сукупності визначається як середнє значення результатів вимірювань по всій множині *об'єктів генералізації* – сукупності результатів вимірювання, яка могла би бути отримана за всіх можливих умов. В загальному випадку ці вимірювання не вважаються строго паралельними. Одним із способів визначення коефіцієнта генералізації є підрахунок відношення дисперсії оцінки в генеральній сукупності до дисперсії очікуваної спостережуваної оцінки.

Розглянемо для прикладу найбільш простий випадок дослідження однофасетних дизайнів, коли єдиним фасетом є фасет екс-

пертів-оцінювачів. Навіть у цьому найпростішому випадку для Р-дослідження можуть обиратися різні дизайни:

1. Кожен опитуваний оцінюється одним і тим же експертом.

2. Кожен опитуваний оцінюється однією і тією ж групою експертів.

3. Кожен опитуваний оцінюється одним експертом, причому різні опитувані оцінюються різними експертами.

4. Кожен опитуваний оцінюється кількома експертами, причому різні опитувані оцінюються різними групами експертів.

У перших двох дизайнах всі опитувані знаходяться в однакових умовах вимірювання. У цьому випадку кажуть, що фасет *перетинається* з опитуваними. В двох останніх дизайнах опитувані знаходяться в різних умовах вимірювання, і тоді кажуть, що множина умов вимірювання є *вкладеною* у множину опитуваних. Оскільки кожному дизайну відповідають різні дисперсії спостережених оцінок, то й коефіцієнти генералізації для різних дизайнів є різними. Розглянемо перший дизайн. Припустимо, що дослідник передбачає, що дослідження буде довготривалим, і тому, можливо, доведеться у різні періоди часу використовувати різних оцінювачів. Тому дослідник хоче з'ясувати, як у різних сеансах тестування зміна оцінювача може впливати на результат. Для цього йому слід розглядати варіант Г-дослідження, при якому фасет оцінювачів є випадковою множиною з нескінченно великою кількістю осіб. Тут сукупність ймовірних оцінювачів є генеральною сукупністю генералізації.

Розглянемо класичну модель тестової оцінки:

$$X_{pi} = T_{pi} + E_{pi},$$

для p-го опитуваного, оцінюваного i-м оцінювачем. Для p-го опитуваного величина T_{pi} буде змінюватися в залежності від оцінювача, а середнє значення усіх T_{pi} по усіх оцінювачах є його генеральною оцінкою. Позначимо цю генеральну оцінку через μ_p. Позначимо також як μ_i математичне очікування істинних оцінок опитуваних, яких оцінював експерт i. Подібно до класичної теорії, математичне очікування X_{pi} дорівнює математичному очікуванню

T_{pi} істинних оцінок, виставлених експертом i. Позначимо, нарешті, символом μ середнє по генеральних оцінках опитуваних.

Тоді лінійна модель для X_{pi} має вигляд:

$$X_{pi} = \mu + (\mu_p - \mu) + (\mu_i - \mu) + e_{pi},$$

або, в термінах відхилень,

$$X_{pi} - \mu = (\mu_p - \mu) + (\mu_i - \mu) + e_{pi}.$$

Таким чином, відхилення оцінки кожного опитуваного від головного середнього має три компоненти – ефект опитуваного $(\mu_p - \mu)$, ефект оцінювача $(\mu_i - \mu)$, та залишкову похибку e_{pi}. Остання відрізняється від похибки вимірювання E_{pi} тим, що має додатковий компонент, обумовлений тим, що істинні оцінки, присвоювані різними експертами, повністю не скорельовані.

Припустимо, що дослідник проводить Г-дослідження, використовуючи 10 екзаменованих та трьох експертів, кожен з яких виставляє свою оцінку кожному з екзаменованих.

Нагадаємо, що у класичній теорії істинної оцінки надійність набору оцінок може визначатися як відношення дисперсії істинних оцінок до дисперсії спостережених оцінок. У нашому випадку доцільно визначити здатність до генералізації як відношення дисперсії генеральної оцінки до дисперсії спостереженої оцінки $\frac{s_p^2}{s_{X|i}}$ для експерта, який буде працювати в Р-дослідженні, і особа якого наперед не відома. Зокрема, це може бути і особа, якої немає серед тих трьох, які беруть участь у Г-дослідженні.

Не маючи даних для нашого гіпотетичного експерта, ми можемо замінити дисперсію в знаменнику її оцінкою – середнім значенням дисперсії спостережених оцінок для всіх експертів генеральної сукупності, яка дорівнює $s_p^2 + s_e^2$. В свою чергу, для підрахунку цієї величини можна використовувати дані від трьох експертів, які беруть участь у Г-дослідженні. Остаточно отримаємо коефіцієнт генералізації

$$\rho_{i^*}^2 = \frac{s_p^2}{s_p^2 + s_e^2}.$$

Зірочка в позначенні коефіцієнта генералізації означає, що цей коефіцієнт годиться для Р-дослідження з умовами вимірювання, що перетинаються з множиною опитуваних. Зауважимо, що ми використали дані від трьох експертів для оцінки дисперсії спостережених оцінок, маючи на увазі, що ці три експерти представляють ту ж саму генеральну сукупність, до якої належатиме й гіпотетичний експерт, який буде брати участь у Р-дослідженні.

Коефіцієнт генералізації може бути оцінений методами двофакторного дисперсійного аналізу ANOVA. Факторами, в термінах ANOVA, є опитувані та експерти. Кожен опитуваний представляє один *рівень* фактора опитуваних, кожен експерт – один *рівень* фактора експертів.

Таблиця 5.3. Оцінки 10 опитуваних, виставлені 3 експертами

Опитуваний	Експерт			Середнє X_{pi}
	1	2	3	
1	2	3	2	2,33
2	8	5	7	6,66
3	4	2	2	2,66
4	4	3	6	4,33
5	8	5	5	6,00
6	8	5	7	6,66
7	6	4	5	5,00
8	4	3	3	3,33
9	3	2	2	2,33
10	1	2	3	2,00
Середнє	4,8	3,4	4,2	4,13

Обчислення зазвичай проводять за допомогою спеціалізованих статистичних комп'ютерних пакетів, хоча достатньо й табличного процесора. Для контролю правильності використання комп'ютерної програми розглянемо приклад з даними, не вдаючись до пояснення ідей дисперсійного аналізу. Таблиця 5.3 містить початкові дані тестування оцінювання 10 осіб трьома експертами, таблиця 5.4 – розрахункові формули двофакторного дисперсійного

аналізу, таблиця 5.5 – результати обчислень за цими формулами. У таблиці 5.4 використовуються традиційні для дисперсійного аналізу позначення:

SV – джерело дисперсії;
SS – суми квадратів;
df – кількість степенів свободи;
MS – середні значення квадратів;
EMS – очікувані середні квадратів.

Таблиця 5.4. Формули підрахунку для двофакторного ANOVA

SV	***SS***	***df***	***MS***	***EMS***
Опитуваний (*P*)	$n_i \sum_p (X_{pI} - X_{PI})^2$	$(n_p - 1)$	$\frac{SS_p}{n_p - 1}$	$s_e^2 + n_i s_p^2$
Експерт (*I*)	$n_p \sum_i (X_{Pi} - X_{PI})^2$	$(n_i - 1)$	$\frac{SS_i}{n_i - 1}$	$s_e^2 + n_p s_i^2$
Залишкові компоненти (*R*)	$\sum_i \sum_p (X_{pi} - X_{PI})^2 - SS_p - SS_i$	$(n_p - 1) \times (n_i - 1)$	$\frac{SS_r}{(n_p - 1)(n_i - 1)}$	s_e^2
де $X_{pI} = \sum_i \frac{X_{pi}}{n_i}$, $X_{Pi} = \sum_p \frac{X_{pi}}{n_p}$, $X_{PI} = \sum_i \sum_p \frac{X_{pi}}{n_i n_p}$				

Таблиця 5.5. Результати підрахунків для ANOVA

SV	***SS***	***df***	***MS***	***EMS***
Опитуваний (*P*)	92,794	9	10,310	$s_e^2 + n_i s_p^2$
Експерт (*I*)	9,866	2	4,933	$s_e^2 + n_p s_i^2$
Залишкові компоненти (*R*)	18,780	18	1,043	s_e^2

Процедура оцінювання значення ρ_{i*}^2 полягає у підрахунку вибіркових оцінок компонентів дисперсії s_p^2 та s_e^2 з використанням зважених комбінацій величин *MS* з таблиці 5.5:

$$s_p^2 = \frac{(MS_p - MS_r)}{n_i}, \quad s_e^2 = MS_r = 1{,}043.$$

Для нашого прикладу отримаємо:

$$s_p^2 = \frac{(10{,}310 - 1{,}043)}{3} = 3{,}089.$$

Тоді, оскільки

$$\rho_{i^*}^2 = \frac{s_p^2}{s_p^2 + s_e^2},$$

то, підставивши значення, отримаємо $\rho_{i^*}^2 = 0{,}75$.

Альтернативна формула для підрахунку оцінки коефіцієнта генералізації:

$$\rho_{i^*}^2 = \frac{MS_p - MS_r}{MS_p + (n_i - 1)MS_r}.$$

Підставивши у цю формулу відповідні значення, отримаємо той самий результат:

$$\rho_{i^*}^2 = \frac{10{,}310 - 1{,}043}{10{,}310 + (3 - 1)1{,}043} = 0{,}75.$$

Отримана величина коефіцієнта генералізації є значущою. Зауважимо, що у нашому прикладі дослідник проводить генералізацію, передбачаючи залучення до оцінювання деякого гіпотетичного експерта, якого немає серед трьох тих, які брали участь у Г-дослідженні. Якби передбачалося, що в Р-дослідженні братиме участь один з цих трьох експертів, то формула для підрахунку оцінки коефіцієнта генералізації мала б інший вигляд, і ми отримали б для нашого прикладу значення 0,90.

Ми довели до кінця розгляд процедури генералізації лише одного виду Р-дослідження, маючи на меті передусім показати необхідний об’єм роботи. Практичне застосування теорії генералізації вимагає ґрунтовного вивчення спеціальної літератури.

6. МЕТОДИ ДОСЛІДЖЕННЯ ВАЛІДНОСТІ

Змістова валідність. Як ми вже зазначали, перевірка змістової валідності тестових завдань є неформалізованою процедурою, яка виконується експертами з цільової області вимірювання на найбільш ранніх етапах конструювання тесту. У процесі змістової валідизації можуть брати участь як автори так і сторонні особи. Загалом процес валідизації складається з наступних кроків:

1. Конкретизація цільової області.
2. Відбір компетентної групи експертів.
3. Забезпечення методики з'ясування відповідності завдань цільовій області вимірювання.
4. Збір даних та підведення підсумків щодо змістової валідності завдань та тесту в цілому.

Перш за все, дослідник повинен визначитися, чи повинен перелік характеристик поведінки представників цільової популяції, який представляє вимірювану якість, бути зваженим, чи ці характеристики повинні вважатися рівноважливими. Оскільки змістова валідність найбільш часто перевіряється у тестах навчальних досягнень, визначатимемо далі цей перелік характеристик як перелік навчальних цілей. Якщо буде прийнято рішення про необхідність зважувати цілі, то одним із прийнятних методів є присвоювання їм балів, скажімо, за п'ятибальною шкалою. Бажано, щоб ранжування цілей відбувалося якомога більшою кількістю експертів, в тому числі представників органів управління освіти. Потрібно конкретизувати підхід до визначення важливості цілі: це може бути як час, що відводиться на її досягнення, так і її роль у процесі засвоєння цільової області.

Для з'ясування відповідності завдань цільовій області можна запропонувати експертам спочатку самим відповісти на завдання, виступаючи у ролі екзаменованих. Рекомендується порівняти кожне завдання із списком навчальних цілей і записати результат порівняння у заздалегідь визначеній стандартній формі. Порівняння з кожною окремою ціллю може проводитися як у дихотомічній формі (відповідає-не відповідає), так і за числовою шкалою, ска-

жімо, п'ятибальною, де 1 означає погану відповідність, а 5 - максимально добру. Якщо цю роботу виконує кілька експертів, то остаточний результат обчислюється як середнє або медіана оцінок експертів.

Крім відповідності навчальним цілям, потрібно з'ясувати відповідність завдань іншим аспектам, таким як вид пізнавального процесу, рівень складності пізнавального процесу, форма стимулу (завдання), способи одержання та представлення потрібної відповіді. Розглянемо приклад (Крокер, Алгіна).

Нехай для дітей певного віку визначені такі дві навчальні цілі з математики:

А. Додавання будь-яких двох додатних цілих чисел, сума яких не перевищує 18.

Б. Віднімання двох цілих чисел, кожне з яких менше 20, і різниця яких є додатним числом.

Нехай пропонується помістити в тест такі 6 завдань:

1. 3 + 5 =
2. 12 – 10 =
3. 8 – 5 =
4. 25 – 16 =
5. 13 + 3 – 8 =
6. У Дмитра було 10 копійок. Він загубив 2 копійки. Скільки копійок у нього залишилось? а) 2; б) 8; в) 10; г) 12

Тут завдання 1 відповідає цілі А, завдання 2 і 3 – цілі Б, завдання 4 не відповідає жодній з цілей. Завдання 5 відповідає обом цілям, але вимагає дещо вищого рівня дій, ніж визначено цілями, оскільки для його виконання потрібна комбінація умінь, визначених у цілях. Завдання 6 відповідає цільовій області, але вимагає уміння читати, чого, в принципі, може й не передбачатися. Крім того, потрібно з'ясувати, чи запис операцій у рядок відповідає тому, до якого звикли діти (вони могли на уроках використовувати вертикальний формат запису).

Спосіб пред'явлення завдань та отримання відповідей (усно, письмово, на комп'ютері), взагалі кажучи, впливає на зв'язок між завданнями та навчальними цілями, особливо це стосується мовних тестів.

Хоча проблема змістової валідизації є швидше якісною, ніж кількісною, тим не менше при підсумовуванні результатів, отриманих для окремих завдань, можуть використовуватися певні кількісні характеристики. Такими характеристиками, зокрема, можуть бути:

1) відсоток завдань, які відповідають цілям;

2) відсоток завдань, які відповідають цілям з високою вагою важливості;

3) кореляції між вагами важливості цілей і кількостями завдань, що їм відповідають;

4) відсоток цілей, які не досягаються у жодному з завдань;

5) інші показники, такі, як, наприклад, показник конгруентності завдань і цілей, запропонований Хамблетоном і Ровінеллі.

Зрозуміло, що використання різних показників може привести до різних результатів. Перші два показники з перерахованих вимагають досить великої кількості завдань (близько 100), щоб їх інтерпретація була значимою. Третій залежить від вагів цілей і варіації кількості завдань. Якщо цілі рівнозначні і кожній з них відповідає однакова кількість завдань, кореляція буде нульовою.

Показник конгруентності, зазначений у п'ятому пункті допомагає виразити відповідність окремого завдання кільком цілям одночасно. В ідеалі цей метод передбачає, що кожне завдання повністю відповідає одній і тільки одній цілі. Спираючись на це припущення, експертам пропонують кожне завдання зіставити з кожною ціллю і присвоїти кожній відповідності 1, якщо завдання відповідає цілі, –1, якщо не відповідає, і 0, якщо не можна з'ясувати відповідність напевне. Тоді показник конгруентності *i*-го завдання *k*-ій цілі може бути обчислений як

$$I_{ik} = \frac{N}{2N-2}(\mu_k - \mu),$$

де N – кількість цілей, μ_k - середня оцінка експертів конгруентності *i*-го завдання *k*-ій цілі, μ_k - середня оцінка експертів *i*-го завдання по всіх цілях. Максимальну оцінку конгруентності 1 можна отримати, якщо завдання було віднесене до однієї і тієї ж цілі усіма експертами. Передбачається, що в результаті оцінювання усіх

завдань кожне завдання повинне мати високий показник конгруентності для запланованої цілі, і низький показник – для незапланованих цілей.

Четверта з перелічених вище характеристик показує, наскільки добре завдання охоплюють цільову область. Цей показник є обернено пропорційним до першого.

У багатьох випадках стандартизованого тестування навчальних досягнень, якщо воно є тестуванням високої відповідальності, бажано забезпечити прийнятну очевидну валідність тестових завдань. З іншого боку, якщо вимірюються психологічні характеристики особистості, краще, якщо опитувані не здогадуються, що саме вимірюється, бо очевидна валідність може зашкодити, оскільки опитувані можуть намагатися відповідати на питання у відповідності з тим, якими би вони хотіли бути чи здаватися для оточуючих, а не якими вони є насправді.

Критеріальна валідність. Як ми зазначали раніше, потрібно розрізняти два види критеріальної валідності – прогностичну та конкурентну. Дослідження кожного з цих видів критеріальної валідності має свої особливості.

У загальному випадку для дослідження критеріальної валідності потрібно виконати наступні кроки:

1) знайти відповідний критерій та спосіб його вимірювання;
2) сформувати придатну вибірку екзаменованих, репрезентативну щодо цільової популяції;
3) пред'явити тест екзаменованим та отримати результати;
4) коли дані по критерію будуть доступні, визначити міру виконання критерію для кожного екзаменованого;
5) визначити кореляцію між результатами тестування та критеріального вимірювання.

Відмінність між дослідженням прогностичної та конкурентної валідності проявляється на четвертому кроці цієї послідовності.

Найбільші проблеми при дослідженні критеріальної валідності вимірювання виникають у зв'язку з ідентифікацією критерію. Також далі ми торкнемося таких проблем, як недостатній об'єм вибірки, контамінація критерію, обмеження діапазону, ненадій-

ність *предиктора* – вимірювання, для якого оцінюється критеріальна валідність.

За Торндайком, можна поділити міри критерію на безпосередні, проміжні та остаточні. *Безпосередні* міри критерію є легко доступними і простими для вимірювання, наприклад, оцінка з деякого навчального курсу, експертна оцінка спостерігача за роботою медичної сестри під час виконання нею ін'єкції, час, необхідний секретарю для того, щоб підготувати та роздрукувати стандартний лист. Такі критерії часто бувають недостатньо повними. *Остаточні* ж критерії, навпаки, володіють характеристиками повноти, але можуть бути складними з точки зору операційного визначення та вимірювання. Прикладами остаточних критеріїв є «педагогічна компетентність», «ефективність роботи вчителя», «незалежність у діях». Тобто остаточні критерії фактично є конструктами.

Припустимо, що ми хочемо валідизувати тест для передбачення ефективності роботи в клас майбутніх вчителів. Тоді відповідний критерій повинен би визначатися на повторюваних спостереженнях за роботою вчителів протягом достатньо довгого (скажімо, 5 років) періоду після закінчення ними ВНЗ. Як бачимо, подібний критерій є дуже складним з практичної точки зору. Тоді ми були б змушені вдатися до *проміжного* критерію, на основі оцінки діяльності студентів при проходженні ними виробничої практики. А безпосереднім критерієм для цього випадку могли би бути оцінки за підготовку студентами планів уроків. Із наведеного прикладу випливає потреба у компромісному виборі критерію. З одного боку, ми хотіли би використовувати найбільш надійний та інформативний остаточний критерій, але за браком часу та інших ресурсів (людських, матеріальних) змушені вдовольнятися безпосередніми або проміжними критеріями.

Розмір вибірки критично впливає на точність коефіцієнтів валідності, якими в критеріальному дослідженні валідності вважаються коефіцієнти кореляції між результатами тесту і критеріальною мірою. Дослідження показують, що предиктор, який є достатньо валідним для популяції, на вибірці 30-50 екзаменованих, є валідним на рівні лише 25-30%. Для невеликих навчальних закладів, які хочуть провести дослідження критеріальної валідності власних тестів, але не володіють достатньо великими вибірками учнів чи студентів, виходом із ситуації може бути дослідження

придатності близьких критеріїв, які вже є у розпорядженні розробників тесту.

Ефект *контамінації критерію* означає суб'єктивний вплив результатів тестування на результати вимірювання критерію. Наприклад, викладачі університету, знаючи про високі показники тестів ЗНО деяких студентів, можуть, свідомо чи несвідомо, завищувати їм оцінки при підсумковій атестації за курс, які передбачається використовувати як критерій. Це призводитиме до штучного збільшення кореляції між тестом і критерієм. З іншого боку, викладачі, які знають про низькі оцінки студентів за тести ЗНО, можуть докласти додаткових зусиль, щоб навчити цих студентів, в результаті чого зв'язок між тестом та критерієм зменшиться. В кожному випадку ефект контамінації критерію бажано усунути у той чи інший спосіб.

Обмеження діапазону означає зменшення дисперсії результатів у вибірках. Подібне відбувається в двох випадках. Перший випадок виникає в ситуаціях, коли тест використовується для відбору ще до того, коли його валідність з'ясована. Це суперечить теорії, але на практиці все ж відбувається. Яскравим прикладом є визначення прогностичної валідності тесту ЗНО з дисципліни, коли критерієм обирається результат навчання студента на першому курсі (скажімо, середній бал з основної дисципліни чи групи дисциплін). Оскільки до університету потрапляють не всі, хто проходив тестування, а здебільшого особи з кращими результатами, то відбувається природне звуження вибірки, і вона перестає бути репрезентативною щодо цільової популяції випускників шкіл. Припустимо, що деяка спеціальність в університеті настільки популярна, що на навчання змогли потрапити лише ті вступники, які отримали 200 балів з відповідного тесту ЗНО. Тоді дисперсія тестових оцінок у вибірці першокурсників взагалі дорівнює нулю і, відповідно, спостерігатиметься нульова кореляція між результатами тестування і оцінками за перший курс. Зменшення дисперсії по предиктору також спостерігатиметься, якщо відбір виконується за допомогою деякої іншої змінної, наприклад, середнього бала атестату випускника школи, оскільки існує кореляція між

У іншому випадку обмеження діапазону відбувається, якщо міра предиктора чи критерію є надто низькою, або навпаки, надто високою.

Існують методи оцінювання коефіцієнтів валідності на основі обмежених щодо даних предиктора або критерію груп. Однак ці методи вимагають припущень, які можуть виявитися ненадійними або їх неможливо перевірити на практиці. Одне з таких припущень полягає у тому, що лінійна регресія критерію по предиктору є однією і тією ж для всіх значень предиктора, тобто для групи відібраних осіб і для групи осіб, які не пройшли відбору. Інше припущення полягає в тому, що дисперсія умовних розподілів критерію для різних значень предиктора є однаковою. В усякому разі, щоб уникнути ефекту обмеження діапазону, і не покладатися на виконання наведених припущень, краще проводити процедуру валідизації тесту ще до того, коли буде здійснено відбір.

Насамкінець зауважимо, що максимально можлива кореляція між предиктором і критерієм прямо залежить від *надійності* обох. Поняття надійності розглядалося нами раніше. Нагадаємо, що надійність вимірювання означає стійкість результатів при повторному тестуванні, а також внутрішню узгодженість (корельованість) між тестовими завданнями. Очевидно, слід добиватися одночасно високої надійності як предиктора, так і критерію. Але слід пам'ятати, що висока внутрішня узгодженість тесту не вкладає в критерій додаткової дисперсії, зменшуючи тим самим валідність. Наприклад, якщо між відповідями на два завдання тесту існує максимальна кореляція (тобто всі екзаменовані відповідають на кожне з цих завдань однаково успішно або не успішно, то одне з завдань не додає дисперсії до критерію. Щоправда, одне з завдань є тут просто лишнім, його слід видалити з тесту або замінити іншим вже на етапі аналізу результатів польової апробації тесту.

Результати дослідження критеріальної валідності, на відміну від змістової валідності, мають числове вираження. Передусім, це *коефіцієнт валідності*, який у випадку вимірювання предиктора та критерію за неперервними метричними шкалами, має вигляд коефіцієнта кореляції Пірсона, за порядковими шкалами – коефіцієнта кореляції Спірмена або «тау» Кендала. Якщо критерій вимірювався за дихотомічною шкалою (наприклад, «отримав диплом про вищу освіту – не отримав»), а предиктор – за неперервною, то свідченням критеріальної валідності тесту може бути статистично значуща *відмінність між середніми оцінками предиктора*, обчис-

лениими для двох підгруп екзаменованих, утворених відповідно до значень критеріальної оцінки. Нарешті, коли обидві змінні – і предиктор і критерій – виміряні за дихотомічними шкалами (або є сенс привести їх до такого вигляду), коефіцієнт валідності може мати вигляд φ-коефіцієнта кореляції.

Додаткову інформацію дає значення коефіцієнта детермінації (квадрата коефіцієнта кореляції). Як нам вже відомо, це число вказує на ту частину дисперсії залежної змінної, яка привнесена мінливістю незалежної змінної. Якщо, наприклад, кореляція між тестовою оцінкою і деякою мірою реальної діяльності осіб дорівнює 0,6, то значення коефіцієнта детермінації 0,36 вказує на те, що 36% дисперсії результатів реальної діяльності пов'язано з дисперсію предиктора.

У главі 2 ми навчилися передбачати значення критерію за значеннями предиктора за допомогою вибіркового рівняння прямої регресії:

$$y - \bar{Y} = \rho_{XY} \frac{s_Y}{s_X}(x - \bar{X}).$$

Нехай y' – прогнозована оцінка критерію у екзаменованого, котрий має по предиктору оцінку x. Тоді з рівняння прямої регресії знаходимо її значення:

$$y' = \rho_{XY} \frac{s_Y}{s_X}(x - \bar{X}) + \bar{Y}.$$

Стандартна похибка вимірювання, як відомо, для значень критерію обчислюється за формулою:

$$s_{YX} = s_Y \sqrt{1 - \rho_{XY}^2}.$$

Значення стандартної похибки допомагає знайти інтервальну оцінку істинного значення критерію (довірчий інтервал). Зокрема, маючи на увазі, що похибки прогнозу критерію в популяції розподілено за нормальним законом, можна вважати з 68% упевненості,

що істинна оцінка екзаменованого за критерієм потрапить у інтервал $y' \pm 1s_{YX}$, і з 95% упевненості – що вона потрапить в інтервал $y' \pm 2s_{YX}$.

Конструктна валідність. Психологічний конструкт був визначений нами раніше як теоретичне поняття, яке є латентною (прихованою від безпосереднього спостереження) величиною. Прикладами конструктів є «інтелект», «креативність», «інтроверт-екстраверт». Для того, щоб конструкт був корисним, потрібно визначити його на двох рівнях – операційному та семантичному. Операційне визначення конструкту полягає у визначенні процедур, за допомогою яких він може бути виміряним. Але, обмежуючись лише операційним визначенням, ми створюємо конструкт як «річ у собі». Необхідно постулювати зв'язки між цим конструктом та іншими конструктами у межах даної теорії, а також між конструктом та певними критеріями реального світу. У цьому й полягає суть *семантичного* визначення конструкту.

Процес конструктної валідизації вимірювання можна описати як наступні кроки:

1) формулювання теоретично обґрунтованих гіпотез про те, як відмінність опитуваних відносно вимірюваного конструкту пов'язана з їх відмінністю відносно інших конструктів та змінних реального світу;

2) розробка інструментів вимірювання на основі операційного визначення конструкту;

3) збір емпіричних даних, необхідних для дослідження усіх гіпотетичних зв'язків;

4) визначення узгодженості емпіричних даних з теоретичними.

Якщо емпіричні дані про зв'язки даного конструкту з іншими конструктами та змінними реального світу підтверджують постульовані на першому кроці теоретичні зв'язки, то можна зробити висновок про високу конструктну валідність даного вимірювання. Зауважимо, що цього висновку все ж недостатньо для того, щоб вимірювання вважалося взагалі валідним.

Якщо ж емпіричні дані не узгоджуються з гіпотетичними, це може означати що:

1) або конструкт визначено неправильно на теоретичному рівні, тобто теорія є неправильною;

2) або теорія є правильною, але досліджуване вимірювання не є валідним, іншими словами, тест є поганим;

3) або і перше і друге одночасно.

Оцінка конструктної валідності вимагає різнобічної інформації з різних джерел. Зупинимося далі на чотирьох найбільш уживаних методах:

- кореляція між мірою конструкту та мірою іншого конструкту;
- метод контрастних груп;
- факторний аналіз;
- матриця «множинні характеристики-множинні методи»

Класичним прикладом спроби довести валідність вимірювання через *кореляцію з іншим конструктом* можна вважати оцінку кореляції між оцінками з тесту інтелекту та мірами навчальних досягнень школярів або навиків у певній роботі. Природно очікувати сильного зв'язку між цими конструктами. Якби гіпотеза про тісний зв'язок між інтелектом та навчальними досягненнями була неправильною, то поняття інтелекту, по суті, втратило би своє практичне значення. Оскільки значущість конкретних значень коефіцієнтів кореляції залежить від ряду чинників, зокрема, об'ємів вибірок, і повинна щоразу з'ясовуватися за допомогою спеціальних методів перевірки статистичних гіпотез, то дуже бажано, щоб теорія вказувала на орієнтовну очікувану величину цих коефіцієнтів. Також, зважаючи на природну множинність зв'язків між конструктами в теорії і на практиці, бажано досліджувати вклад даного конструкту у мінливість за іншим конструктом не окремо, а наряду з іншими конструктами. Тут на допомогу приходять математичні методи множинної кореляції і регресії.

Суть *методу контрастних груп* полягає у перевірці теоретичних постулатів про те, як вимірюваний конструкт має проявлятися у різних груп, які складають популяцію (наприклад, чоловіки і жінки, розумово відсталі і з нормальним рівнем розвитку, асоціальні та просоціальні). Наприклад, вважається (і підтверджується

численними дослідженнями), що такий конструкт як «швидкість мовлення» краще виражений у дівчат ніж у хлопчиків. Якщо тест на швидкість мовлення не виявить цієї особливості, то фактично напевне він не є валідним. В інших випадках невідповідність розподілу результатів вимірювання по контрастних групах теоретичним уявленням розробника тесту може вказувати й на хибність теорії, або і на те. і на інше. Ця обставина ще раз підкреслює необхідність всебічного аналізу валідності вимірювання з залученням якомога більшої кількості джерел інформації.

Факторний аналіз – один із популярних методів математичної статистики. Не слід плутати цей метод з багатофакторним дисперсійним аналізом. Цікаво, що цей метод виник вперше в психометрії. За допомогою факторного аналізу досліджувалася структура інтелекту людини. Зараз метод широко використовується не тільки в психології, а й у нейрофізіології, соціології, політології, економіці та інших науках. Основні ідеї факторного аналізу були закладені англійським психологом і антропологом Ф. Гальтоном (1822-1911), в розробку методу внесли вклад Спірмен, Терстоун, Кеттел, Пірсон, Хотеллінг. Тут ми познайомимося лише з основною ідеєю факторного аналізу.

Цей метод дозволяє одночасно *виявляти взаємозв'язки* між змінними та *компактно їх описувати*. Досліджуються кореляційні зв'язки між змінними у деякій множині змінних. Ті змінні, які дуже тісно корелюють одна з одною, об'єднуються в одну нову змінну (фактор) і дослідник намагається дати загальне описання цієї змінної. По суті, це один із шляхів, яким, якщо змінні є психологічними характеристиками, може бути виявлена якась латентна характеристика, що може потім бути описаною як новий теоретичний конструкт, наприклад, «логіко-математичне мислення» як один із компонентів більш широкого поняття інтелекту. Також факторний аналіз допомагає виявити у множині змінних найменш істотні, що дозволяє, виключивши їх з розгляду, спростити картину зв'язків без суттєвої втрати глибини її деталізації.

Факторний аналіз може використовуватися для дослідження конструктної валідності вимірювання у двох випадках. У першому випадку розглядається матриця попарних кореляцій між завданнями тесту. Якщо серед завдань спостерігаються групи завдань, що сильно корелюють між собою, ці групи завдань можуть вказувати

на існування факторів, що їх об'єднують. Ці фактори вказують на латентні конструкти. Розробнику тесту залишається лише переконатися, як ці конструкти узгоджуються з тими, які були запропоновані теорією.

У другому випадку розглядається кореляційна матриця для наборів різних тестів або мір. При цьому перевіряється, чи будуть тести або субтести, для яких будується кореляційна матриця, і які створені для вимірювання деякого загального конструкту, ідентифікованими емпірично як спільний фактор.

Розглянемо тепер метод з англійською назвою *Multytraits-Multimethods*, яку можна перекласти українською як «множинні характеристики-множинні методи». В основі застосування цього методу лежить та ідея, що свідченням конструктної валідності вимірювання є достатньо сильна корельованість з іншими мірами цього ж конструкту (*конвергентна валідність*) і відносно менша корельованість між даною мірою конструкту і та мірами інших, проте близьких між собою конструктів (*дискримінантна валідність*).

Розглянемо приклад, отриманий Мошером (Mosher, 1968), та наведений Крокер та Алгіною в [6]. Кожен з трьох різних конструктів – комплекс сексуальної вини (А), комплекс вини ворожості (Б) та етична совість (В), вимірювався трьома різними способами: тестами з завданнями альтернативного вибору (1), з регламентованим вибором (2), та з незавершеними реченнями (3), на вибірці з 62 осіб жіночої статі.

Матриця коефіцієнтів парної кореляції між усіма вимірюваннями представлена у таблиці 6.1. Коефіцієнти надійності, розташовані на головній діагоналі, виділено жирним шрифтом. Коефіцієнти конвергентної валідності підкреслено. Наприклад, коефіцієнт кореляції 0,86 у четвертому рядку вказує на зв'язок між результатами тестування комплексу сексуальної вини за допомогою тестів з завданнями альтернативного вибору та з завданнями регламентованого вибору. Усі інші коефіцієнти таблиці (не взяті в дужки та не підкреслені) є коефіцієнтами дискримінантної валідності. Вони утворюють у матриці так звані трикутники гетеровластивостей.

Оскільки використання цього методу шляхом лише візуального аналізу матриці коефіцієнтів може бути проблематичним

через існування похибки вибірки, описаний метод бажано доповнювати додатковими аналітичними дослідженнями.

Таблиця 6.1. Дані матриці «множинні характеристики-множинні методи»

	Метод 1			**Метод 2**			**Метод 3**		
	А	Б	В	А	Б	В	А	Б	В
Метод 1									
А	**0,95**								
Б	0,28	**0,86**							
В	0,58	0,39	**0,92**						
Метод 2									
А	0,86	0,32	0,57	**0,95**					
Б	0,30	0,90	0,40	0,39	**076**				
В	0,52	0,31	0.86	0,55	0,26	**0,84**			
Метод 3									
А	0,73	0,10	0,43	0,64	0,17	0,37	**0,48**		
Б	0,10	0,63	0,17	0,22	0,67	0,19	0,15	**0,41**	
В	0,35	0,16	0,52	0,31	0,17	0,56	0,41	0,30	**0,58**

Як видно з таблиці, коефіцієнти надійності є в цілому високими (третій метод вимірювання конструктів виявився недостатньо надійним), коефіцієнти конвергентної валідності – в цілому вищі від коефіцієнтів дискримінантної валідності.

У цій главі ми описали деякі основні методи дослідження трьох головних видів валідності вимірювання – змістової, критеріальної та конструктної. У читача не повинно скластися враження, що ці методи є альтернативними. У главі 4 ми показали, наскільки важливим є комплексний підхід у дослідженні валідності вимірювання. Цей комплексний підхід загалом передбачає одночасне використання якомога більшої кількості методів дослідження усіх видів валідності, хоча для різних типів вимірювання акцент може робитися лише на деяких із цих видів. Загалом для планування комплексного дослідження валідності вимірювання слід скористатися схемою побудови інтерпретаційного аргументу та аргументу валідності, описаною у главі 4. Зокрема, дослідження повинне включати й збір даних щодо надійності вимірювання. Оскільки ми

зосередилися у цій главі на трьох основних видах валідності – змістовій, критеріальній та конструктній, то ілюструвати логіку дослідження у цих термінах може наступний приклад. Нехай дослідник вважає, що успішність засвоєння природничих наук в університеті залежить від здібностей студентів до розуміння прочитаних технічних текстів. Тому дослідник розробляє тест на «здатність до розуміння прочитаного технічного матеріалу», який складається з уривків технічних текстів, відібраних з університетських підручників з біології та фізики, і наборів завдань множинного вибору, якими супроводжуються ці уривки тексту. Для того, щоб з'ясувати, наскільки завдання є релевантними текстам, потрібно провести дослідження їх змістової валідності. Також необхідно дослідити критеріальну валідність тесту, щоб з'ясувати, наскільки отримані за нього оцінки узгоджуються з успішністю навчання студентів в університеті. Але високі показники критеріальної валідності все ж не означають, що тестом вимірюється саме здатність до розуміння прочитаного технічного матеріалу, а не якийсь інший конструкт чи рису, наприклад, обізнаність студентів з матеріалом, яку вони могли отримати завдяки навчанню в школах з поглибленим вивченням природничих дисциплін. Адже студенти, які засвоїли інформацію, наведену в текстових уривках тесту, ще до початку навчання в університеті, могли б правильно відповідати на завдання, і не читаючи ці уривки. Таким чином, дослідник повинен показати, що тест вимірює саме те, ради чого він створювався, а не загальну наукову поінформованість студентів чи їх загальні академічні здібності. Це є вже предметом дослідження конструктної валідності даного вимірювання.

7. ПРОГНОЗУВАННЯ ТА КЛАСИФІКАЦІЯ НА ОСНОВІ БАТАРЕЇ ТЕСТІВ

У попередній главі ми описали деякі методи валідизації одиничного вимірювання. Зокрема, йшлося про дослідження прогностичної валідності тесту. Але, як ми зазначали раніше, для кращого передбачення діяльності екзаменованих у майбутньому бажано користуватися даними про рівень вираженості у них більше ніж однієї риси чи конструкту. Кожна риса чи конструкт, які передбачаються використовувати для прогнозування критерію, вимірюється за допомогою «свого» тесту. Разом усі тести складають групу, яку називають *батареєю тестів*. У цій главі ми навчимося досягати максимальної прогностичної валідності батареї тестів. Окремо розглядатимемо два випадки. У першому випадку вважатимемо, що як результати вимірювання за кожним тестом з батареї тестів, так і критерій, є неперервними змінними. Для цього випадку валідизацію батареї тестів здійснюватимемо методами частинної кореляції та *множинної лінійної регресії*. Другий випадок відрізняється від першого тим, що критерієм є дискретна категоріальна змінна, описана як множина класів, наприклад «схильні до депресії – схильні до шизофренії». У цьому випадку процес валідизації батареї тестів полягає у максимально ефективному вирішенні *задачі класифікації*. Вирішити цю задачу можна, зокрема, методами *дискримінантного аналізу*.

Використання частинної кореляції для прогнозування критерію. Розглянемо наступний приклад. Нині в Україні підставою до присвоєння абітурієнту місця у рейтинговому списку вступників до ВНЗ є сума його балів з чотирьох предикторів: балів ЗНО з фахової дисципліни, української мови, додаткової нефахової дисципліни, а також середнього балу атестату про середню освіту. Наразі усі ці чотири предиктори мають однакову «силу» – кожен з них обчислено за шкалою 100-200 балів, і їх сума не є зваженою, тобто є простою сумою чотирьох оцінок. Зрозуміло, що, оскільки метою відбору абітурієнтів до ВНЗ є їх успішне навчання там, потрібно досліджувати сукупну прогностичну валід-

ність цих мір, причому прийнятним критерієм може бути успішність навчання на першому курсі ВНЗ, обчислена, наприклад, як середній бал студентів з дисциплін перших двох екзаменаційних сесій.

Виникає питання: наскільки ефективною є описана схема відбору для досягнення максимальної відповідності вказаному критерію? Що стосується середнього балу атестату, то ця змінна має негативну з точки зору використання як предиктора специфіку: зазвичай ця оцінка у вступників є вищою, ніж бали ЗНО, отже, її вага серед інших предикторів виявляється невиправдано великою. Крім того, однаковий середній бал атестата двох випускників, один з яких закінчив невелику сільську школу, а інший – елітну міську гімназію чи ліцей, навряд чи свідчить про їх однакові фактичні навчальні досягнення. Тобто тут є очевидне порушення фундаментальної вимоги до вимірювання – вимоги справедливості.

Звернімося тепер до оцінок ЗНО. Якщо навіть погодитися з тим, що дисципліни, з яких вступник повинен пройти тести ЗНО для вступу на обрану ним спеціальність, визначені цілком слушно, викликає сумнів правильність того, що вони враховуються у прогнозуванні критерію з рівними вагами. Наприклад, видається слушною гіпотеза про те, що оцінка ЗНО з мови є кращим предиктором для успішності навчання на філологічній спеціальності, ніж на технічній, а оцінка з математики – навпаки. З одного боку, цю гіпотезу легко перевірити, порівнявши відповідні коефіцієнти кореляції предиктора з критерієм. Але вище значення цього коефіцієнта для одного предиктора зовсім не означає, що іншим предиктором потрібно знехтувати, оскільки певна зважена комбінація мір двох предикторів може дати ще кращу кореляцію.

З'ясувати ситуацію допомагають методи множинної лінійної регресії. У главі 3 ми розглянули питання прогнозування значення залежної змінної на основі однієї незалежної змінної. Там же розглядалося поняття частинної кореляції.

Нагадаємо, що частинна кореляція допомагає з'ясувати лінійний зв'язок між двома змінними за умови усунення впливу деякої третьої змінної. Нехай X_1 та X_2 – предиктори, Y – критерій, і відомі вибіркові оцінки коефіцієнтів лінійної парної кореляції: $\rho_{X_1X_2}$, ρ_{X_1Y}, та ρ_{X_2Y}. Нехай потрібно визначити значення частинної кореляції другого предиктора з критерієм за умови фіксованого

значення першого предиктора. Як нам відомо, для цього використовується формула:

$$\rho_{X_2Y|X_1} = \frac{\rho_{X_2Y} - \rho_{X_1Y}\rho_{X_1X_2}}{\sqrt{(1-\rho_{X_1Y}^2)(1-\rho_{X_1X_2}^2)}}.$$

Розглянемо для прикладу дані моніторингового дослідження системи вступу до ВНЗ у 2009 році (журнал «Вісник ТІМО, №4-5 за 2010 рік). За цими даними, коефіцієнт кореляції між результатами тесту ЗНО з української мови та літератури (позначимо як X_1) та середнім балом за навчання на першому курсі (Y) у всій вибірці досліджених осіб (всього 22001 особа) становив 0,479; між результатами ЗНО з математики (X_2) та середнім балом за навчання на першому курсі – 0,306; між результатами тестів ЗНО з математики та української мови і літератури – 0,634. Зауважимо, що наведені значення обчислені за Спірменом (аналог коефіцієнта кореляції Пірсона для порядкових змінних), тобто бралися до уваги не самі бали ЗНО, а їх ранги.

З'ясуємо частинний коефіцієнт кореляції між результатами тестування з математики та середнім балом за 1-й курс, за умови, якщо бал ЗНО з мови та літератури має фіксоване значення:

$$\rho_{X_2Y|X_1} = \frac{0{,}306 - 0{,}479 \times 0{,}634}{\sqrt{(1-0{,}479)(1-0{,}634)}} = 0{,}0053.$$

Отриманий додатний результат свідчить про те, що серед тих, хто отримав однакові оцінки з мови та літератури, більше підходять для навчання у ВНЗ ті, хто мають кращі результати з математики. З іншого боку, значення частинного коефіцієнта кореляції є надто малим, щоб вважати додатковий фактор – оцінку з математики – значущим для відбору.

Також зауважимо, що якщо обидва предиктори мають додатну кореляцію з критерієм і сильно корельовані між собою, то використання другого предиктора може виявитися зайвим з огляду на його затратність.

Частинні кореляції можна узагальнити на більшу кількість змінних. Необхідні обчислення при цьому значно ускладнюються, їх краще виконувати за допомогою спеціальних комп'ютерних програм. Слід пам'ятати, що висновки щодо значущості частинних коефіцієнтів кореляції можна робити лише в тому випадку, коли звичайні парні кореляції між змінними є достатньо надійними.

Ми розглянули використання поняття частинної кореляції у контексті підвищення прогностичної валідності батареї тестів. Але частинну кореляцію можна використовувати також і для дослідження критеріальної валідності. Якщо зв'язки між конструктами, передбачені теорією, підтверджуються даними частинних кореляцій, то це є свідченнями на користь правильності теорії та конструктної валідності вимірювань.

Метод множинної лінійної регресії. В главі 2 було показано, як модель лінійної регресії використовується для передбачення результатів одного вимірювання за результатами іншого вимірювання. Пам'ятаючи, що збільшення кількості предикторів може дати більш точний прогноз критерію, ми можемо досліджувати, скільки саме та які з предикторів слід додавати. Для цього використовується модель множинної лінійної регресії. Розглянемо приклад з двома предикторами. У цьому випадку рівняння прямої регресії критерію по предикторах має вигляд:

$$Y' = b_{X_1 Y|X_2} X_1 + b_{X_2 Y|X_1} X_2 + c,$$

де Y' – прогнозована величина критерію, $b_{X_1 Y|X_2}$ та $b_{X_2 Y|X_1}$ – коефіцієнти регресії, c – інтерцепт. Зазвичай усі обчислення проводяться за допомогою комп'ютерного статистичного пакету. Для обчислень вручну можна скористатися формулами:

$$b_{X_1 Y|X_2} = \frac{s_Y(\rho_{YX_1} - \rho_{YX_2}\rho_{X_1X_2})}{s_{X_1}(1 - \rho^2_{X_1X_2})},$$

$$b_{X_2Y|X_1} = \frac{s_Y(\rho_{YX_2} - \rho_{YX_1}\rho_{X_1X_2})}{s_{X_2}(1 - \rho^2_{X_1X_2})},$$

$$c = \bar{Y} - b_{X_1Y|X_2}\bar{X}_1 - b_{X_2Y|X_1}\bar{X}_2.$$

У цих формулах ми використали уведені в главі 2 позначення для вибіркових оцінок середнього, стандартного відхилення, та коефіцієнта кореляції. Розглянемо конкретний приклад.

На фізико-математичному факультеті Ніжинського державного університету імені Миколи Гоголя були зібрані дані про результати тестування з математики (X_1) і української мови та літератури (X_2) вступників 2009 року, прийнятих на навчання, та отримані ними ж середні бали за дві перші сесії, у вигляді оцінок 100-бальної шкали університету. Дані були опрацьовані за допомогою програми *Statistica*. Була отримана така кореляційна матриця для результатів двох тестів та середнього балу за 1 курс навчання (таблиця 7.1):

Таблиця 7.1. Коефіцієнти кореляції для двох предикторів та критерію

	Тест з математики	Тест з укр. мови та літ.	Середній бал за 1 курс
Тест з математики	1		
Тест з укр. мови та літ.	0,555469	1	
Середній бал за 1 курс	0,604300	0,581156	1

Як бачимо, результати навчання на 1 курсі дещо більше корелюють з результатами тестування з математики, ніж з української мови та літератури. Звідси можна зробити висновок, що при вступі тест з математики повинен мати більшу вагу. Але рівняння множинної регресії дає більш точні результати щодо відношення ваг предикторів, оскільки враховує також і взаємний їх зв'язок.

Вага того тесту батареї, чий вплив на загальний прогноз є більш унікальним, є більшою.

У нашому випадку було отримане таке рівняння прямої регресії:

$$Y' = 0{,}407X_1 + 0{,}355X_2 - 99{,}65.$$

При таких вагових коефіцієнтах кореляція між батареєю тестів і критерієм дорівнює близько 0,67 – результат, який вважається дуже добрим для подібних випадків. Тепер можна знайти ваги результатів тестів, які слід присвоїти їм при вступі, нормалізувавши отримані коефіцієнти з рівняння регресії, тобто поділивши кожен з них на їх суму. У нашому випадку отримаємо для тесту з математики приблизно 0,54, для тесту з української мови та літератури – 0,46. Велике значення інтерцепту отримали тому, що критерій вимірювався за шкалою 0..100 балів, а обидва предиктори – за шкалою 100..200 балів.

Зауважимо, що з даного рівняння можна робити висновки про вагові коефіцієнти предикторів для найбільш точного прогнозування, оскільки предиктори вимірювалися в одній і тій же шкалі. Якби це було не так, то потрібно було б скористатися іншою формою рівняння прямої регресії, у якій змінні виражені у вигляді z-оцінок, а відповідні коефіцієнти є стандартизованими коефіцієнтами регресії:

$$z'_Y = \beta_{X_1 Y|X_2} z_{X_1} + \beta_{X_2 Y|X_1} z_{X_2}.$$

Про z-перетворення йшлося у главі 2. Стандартизовані коефіцієнти регресії обчислюються за формулою:

$$\beta_{X_i Y|X_j} = \frac{s_{X_i}}{s_Y} b_{X_i Y|X_j}.$$

Інтерцепт тут завжди дорівнює нулю.

Ми розглянули приклад визначення вагових коефіцієнтів предикторів для випадку прогнозування критерію на основі двох предикторів. У загальному випадку, коли використовується мно-

жина k предикторів, рівняння множинної лінійної регресії має вигляд:

$$Y' = c + b_{YX_1|X_2 \ldots X_k} X_1 + \cdots + b_{YX_k|X_1 \ldots X_{k-1}} X_k.$$

Оцінки коефіцієнтів регресії та інтерцепту у випадку більше ніж двох незалежних змінних знайти вручну дуже складно. Для цього використовуються спеціальні комп'ютерні статистичні пакети, такі як *SPSS* або *Statistica*.

Оцінка точності прогнозування. Досі ми вважали, що набір предикторів визначено наперед, і потрібно знайти лише їх ефективні вагові коефіцієнти. Коли рівняння множинної регресії отримано, виникають два запитання:

1. Наскільки точним є прогноз критерію на основі знайденого рівняння?

2. Чи не можна з множини предикторів підібрати таку їх підмножину, яка б давала найбільш точний прогноз? Можливо, початкова множина предикторів (батарея тестів) є надлишковою, і можна отримати більш ефективний результат передбачення критерію з меншою кількістю незалежних змінних-предикторів. Видаливши з батареї тестів лишні, ми зекономимо на майбутнє людські та матеріальні ресурси, і одночасно підвищимо її прогностичну валідність.

Якби рівняння множинної регресії було отримане на основі результатів усієї цільової популяції, воно мало б скоріш за все дещо інший вигляд, ніж рівняння, отримане на основі вибірки. Мірою, яка допомагає з'ясувати цю відмінність, є так званий *квадрат коефіцієнта крос-валідної кореляції* ρ^2_{cv}. Браун запропонував оцінку цієї величини:

$$R^2_{cv} = \frac{(N-k-3)R_c^4 + R_c^2}{(N-2k-2)R_c^2 + k},$$

де k – кількість предикторів, N – кількість опитуваних,

$$R_c^2 = R^2 - \frac{k(1-R^2)}{N-k-1},$$

$$R_c^4 = R_c^2 - \frac{2k(1-R_c^2)^2}{(N-1)(N-k+1)},$$

В свою чергу, в двох останніх формулах присутня величина R^2, яка для випадку двох предикторів обчислюється за формулою:

$$R^2 = \frac{s_{X_1} b_{X_1 Y|X_2} \rho_{X_1 Y} + s_{X_2} b_{X_2 Y|X_1} \rho_{X_2 Y}}{s_Y}.$$

Зауважимо, що величина R_c^4 не є квадратом величини R_c^2. Якщо якась з цих двох величин виходить від'ємною, потрібно вважати її рівною нулю.

Серед усіх поєднань предикторів слід вибирати те, яке дає найбільшу величину R_{cv}^2. Оскільки ця величина досягає максимуму одночасно з величиною R_c^2, то достатньо аналізувати значення останньої.

Іншою мірою точності прогнозування є *стандартна похибка оцінки Y'*. Для випадку з двома предикторами квадрат цієї величини обчислюється за формулою:

$$SE_{Y'}^2 = \frac{N-1}{N-2} s_Y^2 (1-R^2) \times \left(1 + \frac{1}{N} + \frac{z_1^2 + z_2^2 - 2\rho_{X_1 X_2}^2 z_1 z_2}{(N-1)(1-\rho_{X_1 X_2}^2)}\right),$$

де z_1 та z_2 позначають z-перетворення оцінок X_1 та X_2 відповідно. Стандартна похибка оцінки Y' може використовуватися для визначення довірчого інтервалу прогнозу критерію. Враховуючи, що її розподіл близький до нормального, можна стверджувати, наприклад, з ймовірністю 0,68, тобто з 68% впевненості, що фактичне значення Y потрапить в інтервал $Y' \pm 1SE_{Y'}$.

Ще раз підкреслимо, що множинний регресійний аналіз з більш ніж двома незалежними змінними є складним для ручних обчислень і зазвичай здійснюється за допомогою комп'ютерних статистичних програм. Наведені у цій главі формули допомагають звіряти роботу цих програм для випадку двох предикторів, і є непрямим підтвердженням правильного їх використання з більш великою кількістю предикторів.

Класифікація за допомогою дискримінантного аналізу. Нехай тепер є множина неперервних предикторів, а критерієм є не неперервна змінна, як у попередньому випадку, а деяка множина класів. Приклади:

1. Приймальна комісія деякого ВНЗ бажає віднести кожного абітурієнта на основі його оцінок за тести ЗНО і (або) середнього балу шкільного атестату до однієї з двох популяцій: тих студентів, хто успішно закінчить ВНЗ, або тих, хто його не закінчить. Тут незалежними змінними-предикторами є оцінки ЗНО та середній бал атестату, а залежна змінна представлена двома класами: «закінчить ВНЗ» і «не закінчить ВНЗ».

2. У деякій лікарні пацієнти були віднесені до однієї з двох груп: «хворі на депресію» та «хворі на шизофренію». Створена батарея тестів для вимірювання чотирьох конструктів: ясності свідомості, загальмованості, ворожої підозріливості, депресивної тривожності. Потрібно дослідити придатність цієї батареї тестів до прийняття рішення про віднесення хворого до одного з класів.

Подібні задачі вирішуються методами так званого *дискримінантного аналізу*. Подібно до випадку прогнозування на основі множинної лінійної регресії, дискримінантний аналіз розпочинається з аналізу вибірки тих осіб, для яких правильна класифікація вже відома. Таку вибірку в аналізі даних часто називають навчаючою. Далі на основі навчаючої вибірки виводиться *дискримінантна функція* у вигляді лінійної комбінації змінних-предикторів X_i з ваговими коефіцієнтами a_i. У випадку *n* предикторів дискримінантна функція має вигляд:

$$Y = a_1X_1 + a_2X_2 + \cdots a_nX_n.$$

Вагові коефіцієнти підбираються таким чином, щоб значення *Y* якомога сильніше відрізнялося для представників різних класів критеріальної змінної. Якщо класів критерій представлено лише двома класами, то мірою відмінності між класами природно вважати величину

$$w = \frac{\mu_{Y_1} - \mu_{Y_2}}{s_Y}.$$

У чисельнику тут знаходиться різниця між середніми значеннями функції *Y*, обчисленими окремо для кожної групи осіб, яка утворила клас, у знаменнику – загальна для всієї вибірки осіб дисперсія змінної *Y*. Завдання полягає у підборі таких вагових коефіцієнтів, щоб значення *w* було максимальним.

Спосіб відшукання вагових коефіцієнтів предикторів для дискримінантної функції виходить за межі цієї книги. Для опанування теоретичним матеріалом слід звернутися до літератури з аналізу даних, на практиці ж застосовуються спеціалізовані комп'ютерні програми.

Щойно дискримінантна функція буде побудована, її можна використовувати для класифікації нових осіб на основі їхніх значень предикторів. Для цього потрібно вивести правило віднесення особи до певного класу. Евристичний спосіб виведення такого правила пояснимо на прикладі. Нехай рівняння для дискримінантної функції отримано, і обчислено її значення для кожного представника навчаючої вибірки, яка складається з 10 осіб, по 5 представників кожного з двох класів (таблиця 7.2):

Таблиця 7.2. Значення дискримінантної функції для навчаючої вибірки з 10 осіб.

Клас	**№ особи**	**Значення дискримінантної функції *Y***
I	1	–3,41
	2	–2,62
	3	–3,01
	4	–3,01
	5	–0,84
II	6	–12,70
	7	–14,34
	8	–20,12
	9	–8,42
	10	–0,33

Як видно з таблиці, до першого класу потрапляють особи з загалом більшими значеннями дискримінантної функції. Конструювання рекомендованого правила прогнозування може полягати у

тому, щоб знайти середні значення дискримінантної функції у класах, а потім – середину (середнє арифметичне) між знайденими двома числами. У нашому випадку $\mu_1 = -2{,}0$, $\mu_2 = -11{,}2$, тоді порогове значення

$$c = \frac{\mu_1 + \mu_2}{2} = \frac{-2{,}0 - 11{,}2}{2} = -6{,}6.$$

Отримали правило: якщо значення дискримінантної функції для деякої особи нижче, ніж $-6{,}6$, її слід віднести до II класу, а якщо вище – до I класу.

Як і випадку з множинною регресією, потрібно уміти оцінювати точність прогнозу. Проста оцінка може полягати у ймовірності потрапляння особи до класу, знайдена за даними навчаючої вибірки на основі сконструйованого правила. У нашому випадку усі 5 осіб, що належать до I класу, задовольняють правилу, і одна особа з II класу – не задовольняє. Якби навчаюча вибірка була достатньо великою і репрезентативною, можна було б стверджувати, що ймовірність того, що створене правило коректно віднесе нову особу до I класу, дорівнює 1, а ймовірність того, що правило коректно віднесе особу до II класу – 0,8.

Крім вирішення задачі класифікації, дискримінантна функція може використовуватися для аналізу конструктної валідності предикторів. Якщо всі предиктори виміряні у одній і тій же шкалі, то порівняння коефіцієнтів при предикторах дозволяє зробити висновок про те, який з них має більший вплив на залежну змінну: чим більший коефіцієнт, тим більший вплив. Але у випадку, коли предиктори виміряні у різних шкалах, спочатку слід обчислити *стандартизовані дискримінантні ваги*, і вже по них судити про вплив кожної із незалежних змінних на залежну змінну.

Разом з тим, вважається, що кращим методом інтерпретації дискримінантної функції є аналіз кореляційних коефіцієнтів між кожною із змінних та дискримінантною функцією. Для обчислення кореляції між, наприклад, предиктором X_1 та дискримінантною функцією використовується формула:

$$\rho_{X_1 Y} = a_1 + a_2 \rho_{X_2 X_1} + \cdots + a_n \rho_{X_n X_1}.$$

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%

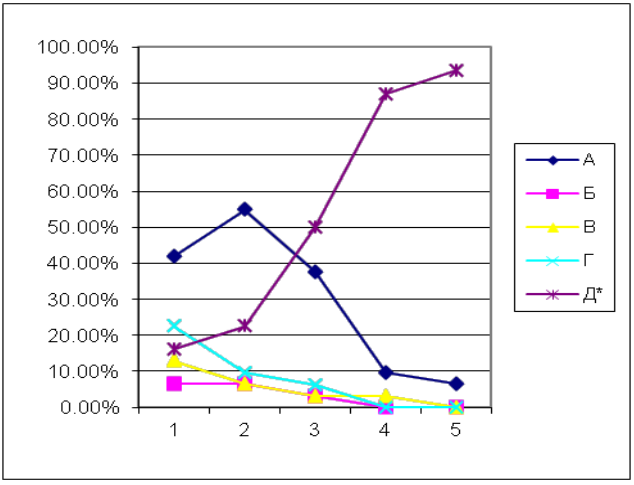


Рис. 8.1. Діаграма розподілу вибору відповідей членами порогових груп

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 корельованість 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 труднiстю 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.

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.

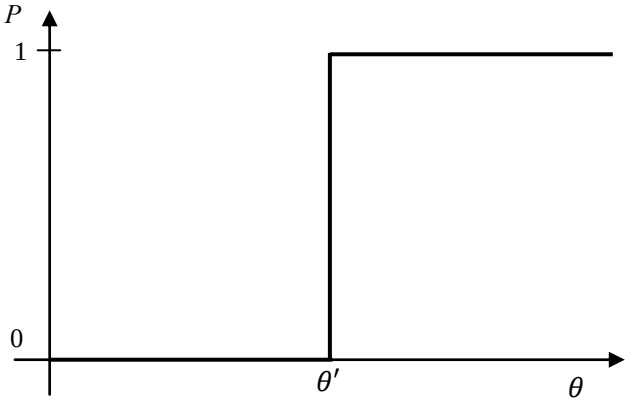


Рис. 9.1. Східчаста функція ICC

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.

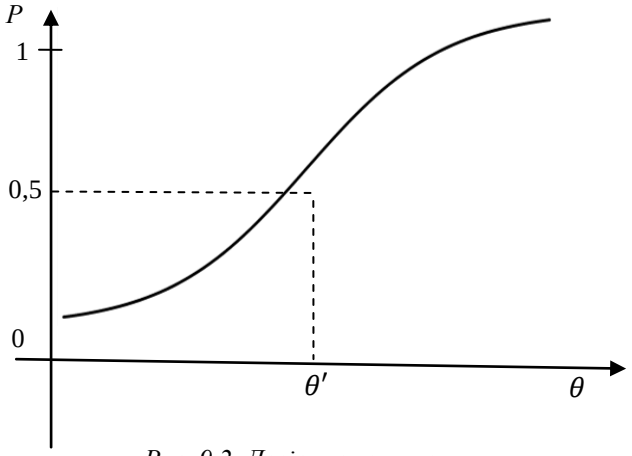


Рис. 9.2. Логістична крива

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 популяії 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).

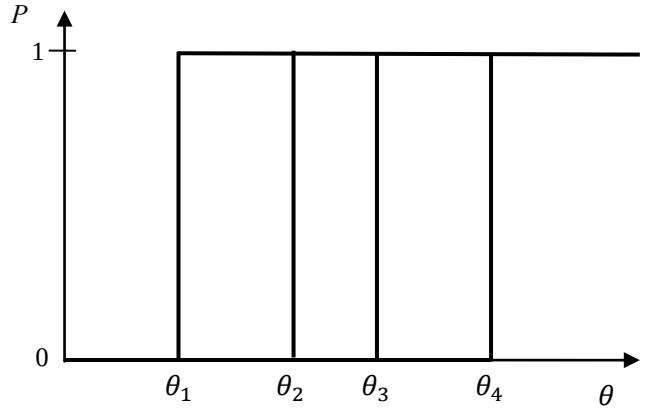


Рис. 9.3. Характеристичні криві 4-х завдань

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: $a_i' > a_i''$.

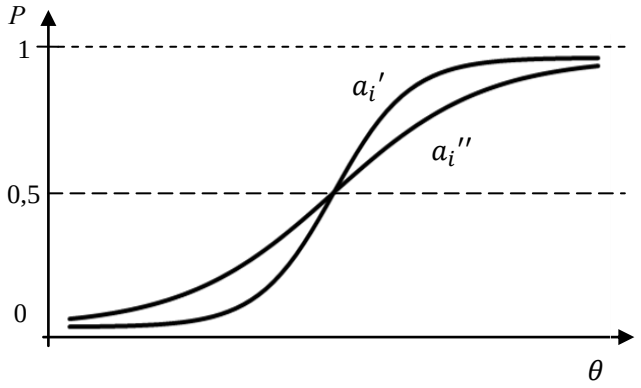


Рис. 9.4. Вплив параметра a_i на форму ІСС

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.

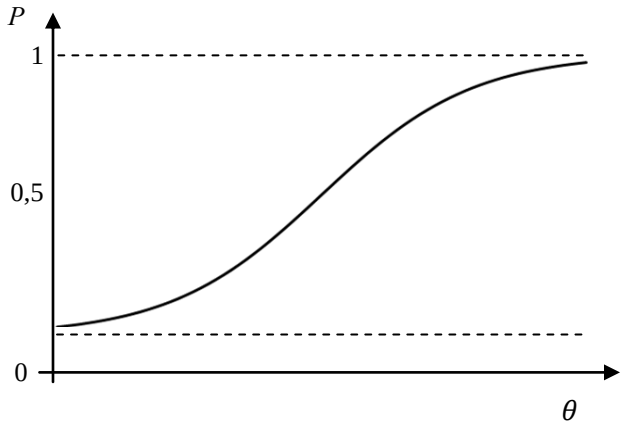


Рис. 9.5. Крива 3PL

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 ІСС 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.

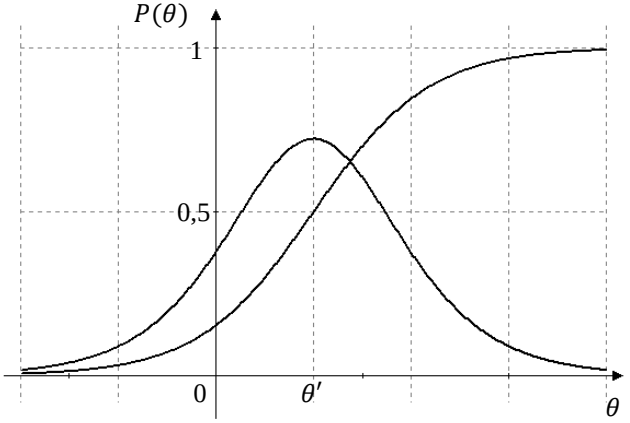


Рис. 9.6. ІСС та інформаційна функція

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.

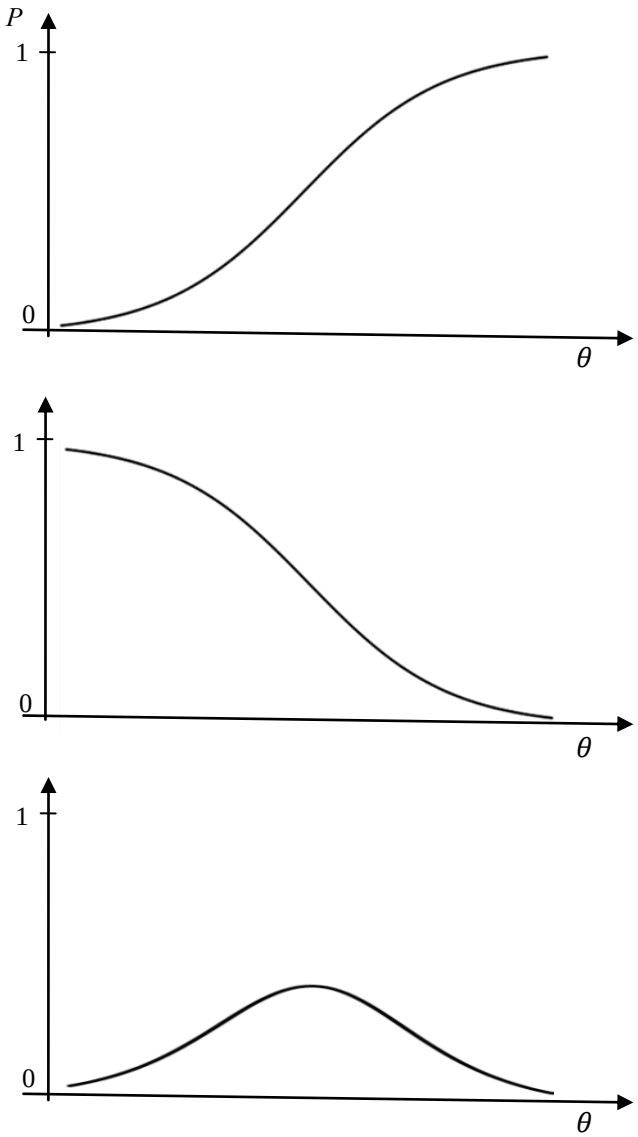


Рис. 9.7. Добуток функцій правильної і неправильної відповідей

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.

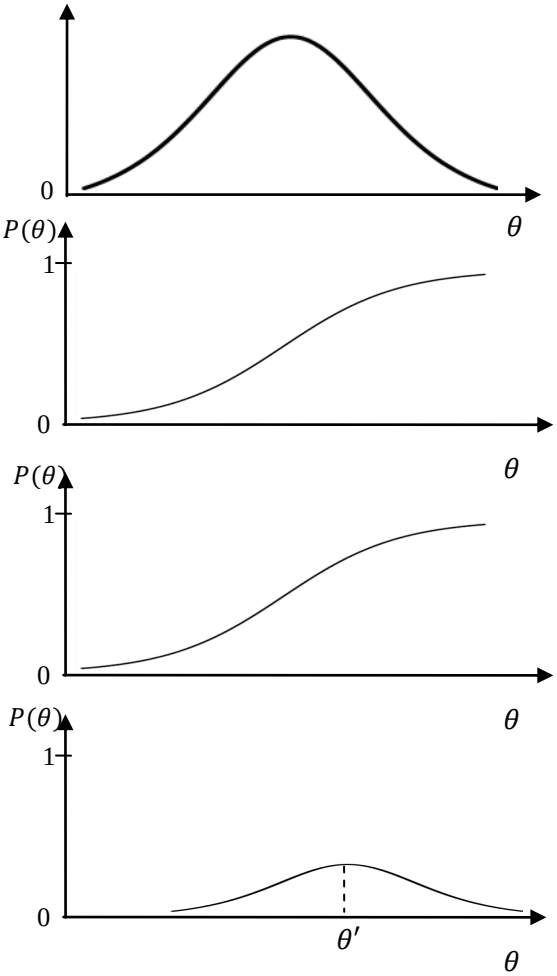


Рис. 9.8. Схематичне представлення Байєсової оцінки

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.

10. ШКАЛЮВАННЯ

Шкалювання (або шкалування) – це процес пов'язування чисел з діями опитуваних під час виконання ними тесту. Результатом шкалювання є поява балів (оцінки) шкали. Часто розрізняють основну і допоміжну шкали, результати обох шкал повідомляють опитуваним. Прикладом допоміжної шкали є процентильні ранги екзаменованих. Інший приклад шкали, яка часто виступає у ролі допоміжної – національна шкала «незадовільно – задовільно – добре – відмінно».

Розробники тесту надають змісту балам шкали для того, щоб ці бали можна було правильно інтерпретувати. Наприклад, якщо 60 балів є середнім значенням для популяції, то бали конкретного опитуваного надають йому інформацію про те, нижчим чи вищим від середнього є його рівень вимірюваної якості. Крім того, повідомлення про бали опитуваних можуть супроводжуватися інформацією про точність вимірювання, а також інформацією про зміст предметної області, який відповідає тому чи іншому балу. Процес шкалювання відіграє також важливу роль при потребі порівняння результатів різних вимірювань.

Для з'ясування прогресу рівня екзаменованого, скажімо, до і після проходження ним навчального курсу, використовується так зване *вертикальне шкалювання*.

Процес шкалювання розпочинається з визначення *первинних балів опитуваних* (англ. *raw scores* – «сирих» балів). У тестах з дихотомічними завданнями (завданнями які оцінюються за дихотомічно шкалою «правильно-неправильно») первинні бали – це кількість правильних відповідей екзаменованого. Первинні бали зазвичай переводяться у бали обраної шкали за допомогою лінійних або нелінійних перетворень.

Шкалювання для одиничного завдання тесту. Слід розрізняти *елементи оцінювання* і *оцінку за завдання*. Нехай, наприклад, кожне завдання оцінюється двома експертами, а оцінкою за завдання є сума балів, виставлених кожним з експертів. Тут оцінка одного експерта є елементом оцінювання, а сума – оцінкою за за-

вдання. Часто елементи оцінювання і оцінка за завдання збігаються, тобто оцінка за завдання містить один елемент оцінювання. Термін «первинні бали» означає функцію від оцінок за завдання.

Співвідношення між елементами оцінювання і оцінкою за завдання може бути відносно складним. Наприклад, при комп'ютерному тестування можуть враховуватися як правильність отриманої відповіді, так і час, витрачений екзаменованим на надання відповіді. У цьому випадку необхідно наперед визначити процедуру поєднання цих двох елементів оцінювання у єдину оцінку за завдання.

Головна відмінність між елементами оцінювання і оцінкою за завдання полягає у тому, що оцінка за завдання вважається операційно незалежною, тоді як між елементами оцінювання існує операційна залежність.

Найбільш поширеними видами оцінки за завдання є наступні.

1. Дихотомічні оцінки. Позначимо через V_i випадкову величину, яка є можливою оцінкою за *i*-те завдання. Для дихотомічної оцінки V_i може набувати одного із двох значень: $V_i = 1$ за правильну відповідь і $V_i = 0$ за неправильну відповідь. Раніше ми вживали термін «дихотомічне завдання» для позначення такого завдання, оцінка за яке є дихотомічною. Зазвичай це завдання множинного вибору з одним варіантом правильної відповіді.

2. Оцінка з поправкою на вгадування. У цьому варіанті оцінки за завдання множинного вибору розрізняються три можливі відповіді: правильна, неправильна і пропущена. Існують різні схеми оцінювання, які стимулюють екзаменованих пропускати відповідь замість того, щоб намагатися її вгадати. Приклад такої схеми:

- $V_i = 1$ за правильну відповідь;
- $V_i = 0$ за пропущену відповідь;
- $V_i = -\frac{1}{A_i - 1}$, де A_i – кількість варіантів відповіді у завданні – за неправильну відповідь.

3. Впорядкована послідовність оцінок. Розглянемо приклад завдання множинного вибору з трьома варіантами відповіді, з яких один варіант – повністю правильна відповідь, другий варіант – частково правильна відповідь, третій варіант – повністю неправи-

льна відповідь. Тоді можна повністю правильну відповідь оцінити у 2 бали, частково правильну – в 1 бал, і неправильну – в 0 балів. Отримали впорядковану множину чисел 2, 1, 0. Така схема оцінювання ще називається *політомічною*, на відміну від дихотомічної, яка насправді є частинним випадком політомічної схеми з послідовністю оцінок 0, 1. Завдання, яке оцінюється за політомічною схемою, можуть називати коротко *політомічним завданням*. Зауважимо, що числа у множині можливих оцінок політомічної схеми не обов'язково мають бути цілими. Можна, наприклад, присвоювати 1 бал за правильну відповідь, 0,5 – балів за частково правильну, і 0 – за неправильну відповідь. Політомічна схема оцінювання виникає й у тому випадку, коли завдання має форму множинного вибору з кількома варіантами цілком правильної відповіді, з можливістю для опитуваного вказати більше ніж один варіант відповіді. Розглянемо, наприклад, завдання:

Столицями держав є міста …
А. Київ
Б. Лондон
В. Нью-Йорк
Г. Сідней

Тут є дві правильні відповіді, інші – неправильні. Розробник тесту може використовувати два підходи у застосуванні політомічної схеми оцінювання такого завдання. Перший підхід полягає у тому, що частково правильні відповіді є допустимими, і вибір неправильної відповіді не є критичним. Наприклад, кожен з правильних варіантів оцінюється в 1 бал, кожен з неправильних – у –1 бал; якщо оцінка за завдання менша від одиниці, вона замінюється на нуль. Тоді, наприклад, екзаменований, який вибрав обидві правильні відповіді і одну неправильну, отримує $1 + 1 - 1 = 1$ бал, екзаменований, який вибрав дві неправильні відповіді у поєднанні з будь-якою кількістю правильних, отримує 0 балів. При такому підході слід контролювати виконання таких вимог: 1) *кожна* неправильна відповідь повинна оцінюватися від'ємним числом; 2) вибір опитуваним *усіх* варіантів відповіді обов'язково повинен оцінюватися як неправильна відповідь, тобто нулем балів; це досягається призначенням за неправильні відповіді такої кількості від'ємних балів, яка не менша за суму балів за правильні відповіді.

Другий підхід, який часом використовується у точних науках, полягає у тому, що вибір хоча б одного неправильного варіанту у поєднанні з правильними варіантами оцінюється нулем балів як повністю неправильна відповідь.

4. Крім розглянутих, часом використовуються схеми з *категоріальними* (невпорядкованими) оцінками, наприклад, «чоловік-жінка», а також з *неперервною* множиною оцінок. Остання є, по суті, впорядкованою дискретною послідовністю з великою кількістю градацій. «Неперервність» тут проявляється швидше у тому, що у групі опитаних значення оцінки за таке завдання можуть не повторюватися, тобто бути різними для різних опитаних. Ці схеми використовуються в освітніх вимірюваннях рідко.

Первинні бали за тест. Первинні бали є функцією на множині балів за окремі завдання, яка поєднує їх в єдину оцінку за тест:

$$X = f(V_1, V_2, \dots, V_n),$$

де n – кількість завдань у тесті. Розглянемо деякі типи первинних оцінок.

*1. Сума балів***.** Найчастіше X визначається як проста сума балів:

$$X = \sum_{i=1}^{n} V_i.$$

Для тестів з дихотомічною схемою оцінювання X збігається з кількістю отриманих правильних відповідей на завдання тесту.

2. Зважена сума. Часом окремим завданням тесту присвоюються вагові коефіцієнти. Тоді

$$X = \sum_{i=1}^{n} w_i V_i,$$

де w_i – ваговий коефіцієнт i-го завдання. Може застосовуватися додаткова вимога, щоб сума всіх вагових коефіцієнтів завдань

тесту дорівнювала одиниці. Вагові коефіцієнти можуть обиратися, наприклад. з огляду на важливість змісту завдань, або для покращення статистичних характеристик тесту.

3. Складна функція. Часом первинні бали визначаються як функція, значно складніша за лінійну комбінацію оцінок за окремі завдання. Прикладом таких функцій є оцінювання за методом максимальної правдоподібності або за модальна Байєсова оцінка, розглянуті нами у главі 9, присвячену основам теорії IRT.

Випадок тестів із змішаними формами завдань. Чи не найчастіше на практиці використовуються тести, у яких присутні завдання різних форм. Наприклад, у тесті можуть зустрічатися як завдання множинного вибору, так і завдання на впорядкування альтернатив. При цьому різні типи завдань можуть оцінюватися різною кількістю балів. Розробники тесту мають у подібному випадку подбати про те, як поєднати бали за окремі завдання в єдину оцінку за тест. Одним із можливих шляхів є приписування вагів завданням пропорційно до очікувано внеску кожної з форм завдань у загальну первинну оцінку. При цьому часто враховують кількість завдань кожного типу, час, який відводиться на відповіді на завдання різних типів, масштаб покриття цими завданнями предметної області.

Інший підхід полягає у тому, щоб присвоювати різні ваги різним формам завдань з огляду на статистичний внесок цих форм у загальну оцінку. У цьому випадку вагові коефіцієнти, присвоювані завданням називають *ефективними вагами*, і відшукують їх за допомогою спеціальних статистичних процедур. Наприклад, є сенс у тому, щоб тим завданням, які сильно корелюють між собою, присвоювати менші вагові коефіцієнти; або ефективні ваги підбираються так, щоб максимізувати надійність тесту.

Перетворення первинної оцінки у оцінку шкали. Первинні оцінки володіють рядом недоліків, через які вони, взагалі кажучи, не можуть вважатися результатами вимірювання. Якщо тетст існує у кількох формах, первинні оцінки не можуть однаково тлумачитися. Наприклад, первинна оцінка 35 балів за простіший тест відповідає більш низькому рівню екзаменованого, ніж ті ж 35 балів за більш складний тест. Цей недолік відсутній у випадку отримання первинної оцінки методами IRT, але й у цьому випадку первин-

на оцінка не дозволяє давати смислову інтерпретацію окремих її значень. Саме смислова інтерпретація оцінок є основним чинником необхідності перетворення первинної оцінки у оцінку обраної шкали. Головними аспектами смислової інтерпретації оцінки є 1)норми, 2) точність оцінювання; 3) тлумачення оцінки у термінах цільової області вимірювання. Далі торкнемося кожного з цих аспектів.

1. Додавання норм до шкали. Процес приєднання нормативної інформації до числових значень шкали розпочинається з того, що тест пред'являється репрезентативній групі осіб з цільової популяції, яка в даному випадку називається *нормативною групою*. На підставі даних про результати тестування нормативної групи отримуються основні статистичні характеристики, такі, як середнє і стандартне відхилення оцінки. Ці характеристики вважаються чинними для усієї цільової популяції, набуваючи тим самим статусу *норм*. Після цього результат тестування будь-якого представника цільової популяції можна порівнювати з нормами для цієї популяції. Наприклад, якщо нормативна середня оцінка становить 50 балів, то результат опитуваного з даної популяції у 60 балів дозволяє вважати, що рівень вимірюваної якості у нього є вищим, ніж середній по популяції.

Головна проблема полягає у визначенні нормативної групи. Чи можна, наприклад, вважати, що група, яка була репрезентативною для популяції випускників шкіл України у 2007 році, репрезентативною для популяції випускників 2012 року?

Припустимо, що нормативна група все ж була визначена і норми отримані. Далі розглянемо можливу трансформацію первинних балів у бали основної шкали з використанням норм.

1.1. Лінійні перетворення. Припустимо, що розробники тесту визначили з певних міркувань, якими мають бути середній бал і стандартне відхилення шкали для створеного тесту. Нехай X – первинна оцінка з конкретним значенням x, $S(x)$ – оцінка шкали, яка відповідатиме значенню x. Тоді формула для лінійного перетворення первинних балів у бали шкали має вигляд:

$$S(x) = \frac{\sigma_S}{\sigma_X} x + \left(\mu_S - \frac{\sigma_S}{\sigma_X}\mu_X\right),$$

де μ_X і σ_X – середнє і стандартне відхилення первинної оцінки, μ_S і σ_S – призначені розробником середнє і стандартне відхилення оцінки шкали.

Переписавши це рівняння у вигляді

$$S(x) - \mu_S = \frac{\sigma_S}{\sigma_X}\left(x - \frac{\sigma_S}{\sigma_X}\mu_X\right),$$

можна побачити, що це, по суті, рівняння прямої регресії S по X з коефіцієнтом кореляції 1 (це означає, що всі точки з координатами (x, s) лежать на прямій лінії).

Розглянемо приклад. Нехай первинні оцінки мають середнє 70 і стандартне відхилення 10. Розробник тесту вирішив, що нормативне середнє шкали повинне дорівнювати 20, а стандартне відхилення – 5. Підставимо значення у наведену вище формулу:

$$S(x) = \frac{5}{10}x - 20 - \frac{5}{10} \times 70,$$

звідки отримаємо формулу для перетворення первинних балів у бали шкали:

$$S(x) = 0{,}5x - 15.$$

Інший шлях застосування лінійного перетворення полягає у тому, що для двох значень первинних оцінок розробник визначає відповідні очікувані бали шкали. Цими двома первинними оцінками зазвичай є мінімальна та максимальна оцінка нормативної групи. Тоді потрібне лінійне перетворення визначається з відомого рівняння прямої, що проходить через дві задані точки:

$$\frac{S(x) - S(x_1)}{x - x_1} = \frac{S(x_2) - S(x_1)}{x_2 - x_1},$$

де x_1 і x_2 – первинні оцінки, $S(x_1)$ і $S(x_2)$ їх відповідники у шкалі, задані розробником. Звідси отримуємо формулу для перетворення будь-якої первинної оцінки у оцінку шкали:

$$S(x) = \frac{S(x_2) - S(x_1)}{x_2 - x_1} x + \left(S(x_1) - \frac{S(x_2) - S(x_1)}{x_2 - x_1} x_1 \right).$$

Наприклад, потрібно діапазон [1, 6] первинних балів, отриманих нормативною групою, перетворити на обрану розробником шкалу [10, 20] балів. Підставивши значення у рівняння прямої, що проходить через дві задані точки (у нашому випадку це точки з координатами (1,10) і (6, 20)), отримаємо:

$$\frac{S(x) - 10}{x - 1} = \frac{20 - 10}{6 - 1} = 2,$$

$$S(x) - 10 = 2x - 2.$$

Остаточно отримуємо рівняння:

$$S(x) = 2x + 8.$$

1.2. *Нелінійні перетворення*. Допустимими є практично будь-які строго монотонні перетворення, оскільки вони зберігають порядок: якщо $x_1 > x_2$, то $f(x_1) > f(x_2)$ для зростаючого перетворення. В окремих випадках розробнику може знадобитись спадне монотонне перетворення: якщо $x_1 > x_2$, то $f(x_1) < f(x_2)$. Це буде фактично означати, що розробник вирішив розглядати замість одного конструкту інший, протилежний йому за змістом, наприклад, замість рівня успішності вимірювати рівень «неуспішності».

Одним із найпростіших нелінійних перетворень є просте округлення балів шкали, отриманих після лінійного перетворення первинних балів.

Ще одне просте перетворення – обрізання шкали. Наприклад, розробник тесту може вирішити, що всі первинні бали, які після лінійного перетворення стали від'ємними числами, слід замінити на шкалі нулем. Округлення і обрізання шкали часто використовуються після лінійного перетворення первинних балів, отриманих методами IRT.

Прикладом більш складного перетворення, яке досить часто використовується, є перетворення первинних балів до шкали з заданою формою розподілу балів. Зокрема, в системі зовнішнього незалежного оцінювання в Україні використовується перетворення, яке називається еквіпроцентильною нормалізацією. Цей метод дозволяє з первинних балів, розподілених довільно, отримати шкалу, бали якої розподілені за законом, близьким до нормального.

Еквіпроцентильну нормалізацію виконують за таким алгоритмом:

1) Визначити розподіл відносних частот первинних балів у нормативній групі.

2) За, необхідністю, цей розподіл згладити одним із відомих методів.

3) Знайти процентильні ранги $Q(x)$ первинних балів.

4) Знайти обернене значення функції нормального розподілу для дробу $Q(x)/100$. Тобто, потрібно знайти таке z, що

$$F_{\text{ст}}(z) = \frac{Q(x)}{100},$$

де $F_{\text{ст}}(z)$ – функція стандартного нормального розподілу $N(0,1)$.

5) Перетворити знайдені значення z до шкали з бажаними значеннями середнього та стандартного відхилення за допомогою лінійного перетворення

$$S(x) = \sigma_S \cdot z + \mu_S$$

(це перетворення є оберненим до z-перетворення $z = \frac{x-\mu_S}{\sigma_S}$.

6) Якщо потрібно, округлити знайдені значення $S(x)$ до цілих чисел і обрізати шкалу.

Зауважимо, що внаслідок цієї процедури ми не отримаємо шкалу з ідеальним нормальним розподілом балів. Тим не менше, асиметрія й ексцес розподілу будуть близькими до нуля, як і для нормального розподілу.

Прикладами нормалізованих шкал є такі відомі шкали:

- *T-шкала* з $\mu_S = 50, \sigma_S = 10$
- *шкала IQ* з $\mu_S = 100, \sigma_S = 15$ (США)
- *шкала станайнів* із цілими значеннями від 1 до 9, середнім 5 і стандартним відхиленням 2.

Перевагою нормалізованих шкал є те, що при інтерпретації їх окремих значень можна використовувати відомі властивості нормального розподілу. Наприклад, оцінка, яка на 1σ перевищує середнє значення, відповідає приблизно 84-му процентильному рангу в нормативній групі. Якщо, скажімо результат тестування IQ у особи становить 115 балів, то можна вважати, що у неї вищий рівень інтелекту (у розумінні розробників тесту), ніж у 84% населення США. Зауважмо, що розподіл первинних балів рідко буває близьким до нормального, зокрема він зазвичай є асиметричним.

Ще одним прикладом часто вживаного нелінійного перетворення є згадані вище процентильні ранги. Часто їх використовують як додаткову шкалу.

2. Додавання до шкали інформації про точність вимірювання. Роздільна здатність шкали повинна відповідати точності, з якою виконувалися вимірювання. Якщо шкала має надто мало поділок, то частина точності буде втрачена. Приклад не з області тестування: якщо один студент отримав за 100-бальою накопичувальною шкалою підсумкову оцінку 91 бал, а інший студент – 99 балів, то обидва студенти за національною 4-бальною шкалою будуть оцінені однаково – на відмінно, тобто інформація про те, хто з них досяг вищого рівня, буде втрачена. З іншого боку, надто велика кількість поділок на шкалі можуть призводити до появи відмінності у оцінках екзаменованих там, де насправді цієї відмінності немає.

У зв'язку з цим були запропоновані різні правила визначення оптимальної роздільної здатності шкали. Ці правила пов'язують кількість поділок шкали з довірчим інтервалом для істинної оцінки екзаменованого. При цьому повинні виконуватися наступні припущення.

1. Шкала є лінійним перетворенням первинних балів.

2. Похибка вимірювання має нормальний розподіл для різних значень істинної оцінки.

3. Стандартна похибка вимірювання є сталою вздовж шкали.

4. Коефіцієнт надійності $\rho_{XX'}$ тесту відомий або знайдена його слушна оцінка.

5. Розмах шкали відповідає шістьом стандартним відхиленням балів шкали (правило «трьох сигм»).

6. Задано довжину довірчого інтервалу h для істинних оцінок екзаменованих.

7. Задано довірчу ймовірність γ для довірчого інтервалу і z_γ – значення стандартного нормального розподілу оцінок, яке формує $100 \cdot \gamma$-відсотковий довірчий інтервал.

Тоді стандартне відхилення шкали може бути обчисленим за формулою:

$$\sigma_S = \frac{h}{z_\gamma \sqrt{1 - \rho_{XX'}}},$$

і кількість поділок шкали можна обчислити, помноживши σ_S на 6 і округливши результат до цілого. Наприклад, нехай $\rho_{XX'} = 0{,}91$ і, згідно з відомим правилом Келлі, покладемо $h = 3$, $\gamma = 0{,}68$, $z_\gamma = 1$. Тоді $\sigma_S = 10$ і, помноживши це значення на 6, отримуємо кількість поділок шкали 60.

Це правило відповідає американському тесту SAT, який має шкалу від 200 до 800 з кроком 10.

Загалом, як нам відомо з глави 5, стандартна похибка вимірювання не є сталою вздовж шкали оцінювання. Для оцінок, отриманих методами класичної теорії, стандартна похибка вимірювання є найбільшою посередині шкали, і меншою на кінцях. Для первинних балів, виміряних методами IRT, картина протилежна. Крім того, нелінійні перетворення первинних балів можуть істотно впливати на розподіл умовних похибок вимірювання. Тому в супровідній документації тесту потрібно вказувати якими є стандартні похибки для різних значень шкали, якщо тільки вона не сконструйована таким чином, що стандартно похибка є сталою для різних її значень. Існують процедури вирівнювання умовних похибок вимірювання Наприклад, якщо первинні бали були отримані як сума балів за правильні відповіді на дихотомічні завдання, з цією метою можна використати нелінійне перетворення

$$g(x) = 0{,}5\left(\arcsin\left(\sqrt{\frac{x}{k+1}}\right) + \arcsin\left(\sqrt{\frac{x+1}{k+1}}\right)\right),$$

де k – кількість завдань у тесті, x – первинна оцінка. Це перетворення використовують для стабілізації дисперсії біноміально розподіленої випадкової величини.

3. Змістове тлумачення оцінок шкали.. Виділяють три процедури, які дозволяють зіставити оцінки шкали із змістом предметної області:

1) відображення (item mapping);
2) прив'язування (scale anchoring);
3) визначення стандартів (standard setting).

Суть *відображення завдань тесту на шкалу (item mapping)* полягає у тому, що для окремих балів шкали вказуються завдання, які за рівнем трудності відповідають цим балам. Для завдань з дихотомічною оцінкою ймовірність правильної відповіді на кожне завдання відображається на шкалу за допомогою процедури логістичної регресії або методами IRT. При цьому задається ймовірність правильної відповіді (*response probability level, RP=level*), зазвичай це значення у межах від 0,5 до 0,8. Нехай, наприклад, прийнято рішення про те, що RP-рівень дорівнює 0,5. Тоді для завдання на шкалі відшукується точка, яка відповідає оцінці за тест тих осіб, половина з яких відповіли на завдання правильно. Особливо легко виконується ця процедура при застосуванні теорії IRT: на осі латентної характеристики відшукується точка, яка відповідає ординаті 0,5 кривої ICC. Якщо задано вищий RP-рівень, наприклад, 0,8), то точка на шкалі, що відповідає завданню, буде знаходитися далі у бік зростання рівня латентної характеристики. Таким чином, вибір рівня RP істотно впливає на змістову інтерпретацію оцінок шкали. Застосовують також додаткові умови. Наприклад, для відображення можуть обиратися лише ті завдання тесту, які володіють високою роздільною здатністю.

У тестах NAEP (National Assessment of Educational Progress, довготривала програма оцінювання прогресу в освіті США) 1996

року з природничих наук для учнів 4-го класу застосовувалося відображення з RP-рівнем 0,74.

Зокрема, наводяться такі дані про результати відображення:

- балу 201 (за шкалою 200-300) відповідає завдання, яке відображає уміння показати на карті Атлантичний та Тихий океани;
- 140 балам відповідає розуміння, як отримують кисень;
- 94 балам відповідає уміння ідентифікувати інструмент, призначений для спостереження за зірками.

Як бачимо, інформація відображення дещо відрізняється від формулювання самих завдань: вона подається у термінах знань і умінь. Тим не менше, тлумачення балів шкали не виходить тут за межі тієї інформації, яка міститься в окремих завданнях. Очевидно, більш інформативною була би більш узагальнена інформація про знання та уміння екзаменованих, які відповідають тій чи іншій оцінці шкали.

Саме у цьому напрямку ідея відображення розвивається далі у процедурі *прив'язування завдань до шкали (scale anchoring)*. Ця процедура полягає у визначенні в загальних термінах того, що саме екзаменовані, які отримали певні оцінки, знають і уміють. На першому кроці прив'язування відбувається розглянута вище процедура відображення. Потім обирається множина точок на шкалі, зазвичай з рівними інтервалами між ними, або множина процентильних рангів, які підлягають тлумаченню у термінах знань і умінь екзаменованих. Наприклад, обираються процентильні ранги із значеннями 10, 25, 50, 75, 90. Ті завдання, відображення яких на шкалу відповідають цим точкам або знаходяться поблизу них, аналізуються експертами для узагальненого описання відповідних знань і умінь екзаменованих. Також беруться до уваги всі ті завдання, відображення яких знаходяться нижче на шкалі від обраних для тлумачення оцінок.

Процедура *визначення стандартів* використовується зазвичай у тестах на професійну придатність, при ліцензуванні чи сертифікації професійної діяльності. Наприклад, визначається мінімум знань і умінь, який повинен продемонструвати екзаменований, щоб отримати ліцензію. Для цього мінімуму відшукується відповідна точка на шкалі оцінок. Для тестів навчальних досягнень визначаються певні рівні, виражені у термінах з певним змістовим

наповненням, наприклад, «задовільно-добре-відмінно», і після цього так само відшукуються точки шкали оцінок, які відповідають переходам від одного рівня до іншого. Як бачимо, процедура визначення стандартів є ніби оберненою для процедури scale anchoring: там спочатку обираються точки на шкалі, а потім ці точки інтерпретуються; тут, навпаки, обираються значимі, на погляд експертів, рівні інтерпретації, а потім для них відшукують відповідні точки на шкалі оцінок.

11. ПОРІВНЯННЯ РЕЗУЛЬТАТІВ ВИМІРЮВАНЬ

Під *порівнянням* результатів різних вимірювань будемо розуміти ту частину теорії освітніх вимірювань і загалом психометрії, яка в англомовній літературі позначається терміном *linking* (зв'язування, з'єднування), і який означає трансформацію оцінок шкали одного тесту в оцінки шкали іншого тесту, або трансформацію оцінок двох тестів у єдину шкалу.

Методи порівняння результатів різних вимірювань можна поділити на три групи: передбачення, вирівнювання шкал, і пряме прирівнювання.

Передбачення (англ. *predicting*) застосовується у ситуаціях, коли необхідно передбачити оцінку опитуваного з деякого тесту на основі спостереженої оцінки з іншого тесту або батареї тестів, чи інших джерел інформації. Цей вид порівняння розглядався нами в главах 2 (для ситуації з одним предиктором) і 7 (для ситуації з кількома предикторами), тому у цьому розділі ми не розглядатимемо його, згадавши його тут лише для повноти викладу. Прикладом ситуації, коли використовується передбачення, є дослідження прогностичної валідності тесту. Проте це далеко не єдина ситуація, яка вимагає використання методів прогнозування.

Вирівнювання шкал (англ. *scale aligning*) означає трансформацію оцінок двох різних тестів у єдину шкалу. Прикладом вирівнювання шкал є переведення оцінок двох тестів з певної дисципліни у системі зовнішнього незалежного оцінювання, які пред'являються різним групам випускників шкіл у двох сесіях тестування, в нормалізовану шкалу методом еквіпроцентильної нормалізації.

Ще одним прикладом вирівнювання шкал, щоправда не пов'язаним з тестуванням, є приведення оцінок накопичувальної 100-бальної шкали в українських університетах у буквену шкалу ECTS (A, B, C, D, E, Fx, F), за тієї умови, що будуть дотримані рекомендації щодо частотного розподілу оцінок. Наприклад, передбачається, що оцінку A повинні отримати приблизно 10 відсотків кращих студентів і т.д. У такому випадку, за умови, що множини студентів у двох різних університетах, які вивчали дану дис-

ципліну, належать одній і тій же популяції з однаковим розподілом рівня успішності з цієї дисципліни, оцінку, отриману в одному університеті, можна перезарахувати в іншому університеті. Якщо ж оцінки за 100-бальною шкалою виставлялися без будь-якого дотримання розподілу частот, а потім механічно переводилися в шкалу ECTS (якщо отримано від 90 до 100, то A, і т.д.), то тоді вже немає ніякої змоги перезарахувати оцінку в іншому університеті, не ризикуючи порушити принцип справедливості оцінювання.

Зауважимо, що останнім часом Європейська комісія, розуміючи особливості національних систем оцінювання, пропонує відмовитися від переведення всіх оцінок в єдину «ідеальну» шкалу ECTS, рекомендуючи натомість використовувати пряме прирівнювання розподілів оцінок, отриманих достатньо великими групами студентів в обох університетах. *Прирівнювання* (англ. *equating*) об'єднує в собі методи відшукання взаємно-однозначної відповідності між шкалами оцінок двох різних тестів, тобто трансформації оцінок шкали одного вимірювання у оцінки шкали іншого вимірювання і навпаки. Саме властивість зворотності трансформації є причиною виділення цих методів в окрему групу.

В усіх випадках порівняння різних вимірювань повинні виконуватися певні додаткові припущення, які дозволяють порівнювати або інструменти вимірювання, або популяції, для яких використовуються ці інструменти. Якщо вимірювати зріст марсіанина у марсіанських одиницях довжини, а жителя Венери – у венеріанських одиницях, то не буде жодної змоги сказати, хто з них вищий, якщо не буде додаткової інформації, яка б дозволила порівняти або одиниці вимірювання (інструменти), або зріст жителів цих двох планет. Так само, пам'ятаючи про відмінність вимірювань у сфері психіки від фізичних вимірювань, доводиться визнати, що не можна порівняти результати двох різних тестувань, проведених на двох різних популяціях, не маючи додаткової «якірної» (англ. anchor) інформації, яка є своєрідним містком або між тестами, або між популяціями. Розглянемо детальніше методи вирівнювання та прямого прирівнювання.

Методи вирівнювання шкал. Цю групу методів можна поділити на дві підгрупи.

Перша підгрупа стосується ситуацій, коли потрібно порівняти результати двох тестувань, якими вимірюються різні конструк-

ти. Тут теж можливі два випадки. У першому випадку вважається, що групи, яким пред'являлися тести, належать до однієї популяції. У другому випадку вважається, що тести пред'являлися групам з різних популяцій. Перший випадок іноді називають шкалюванням батареї (battery scaling). Те, що дві групи належать одній популяції, означає, що це або одна і та ж група, або групи еквівалентні з точки зору розподілу у них рівнів вимірюваної якості. Оскільки у цьому випадку результати кожного з тестувань можуть бути поширені на всю популяцію, то достатньо виконати трансформацію розподілу оцінок кожного з тестів (точніше, емпіричної функції розподілу) у спільну шкалу. Наприклад, батарея тестів SAT I (тесту здібностей до навчання), яка складається з двох субтестів – вербального та логіко-математичного, була приведена до єдиної шкали у 1990 році. Результат такого вирівнювання дозволяє стверджувати, що екзаменований, оцінка якого за математичний субтест є вищою за його оцінку з вербального субтесту, має рівень математичних здібностей вищий, ніж вербальних здібностей.

У другому випадку, коли групи екзаменованих, у яких вимірювалися різні якості, є різними, необхідно залучати додаткову, якірну інформацію. Відповідні методи називають *якірним шкалюванням* (*anchor scaling*). Зауважимо, що у деяких випадках тести, які пред'являються різним групам, є вимушено різними. Наприклад, при тестуванні з іноземної мови в системі ЗНО одні випускники складають тест з англійської, інші – з французької, залежно від того, яку мову вони вивчали. Якірне вимірювання – це спільне для обох груп вимірювання. Воно має бути достатньо сильно корельованим з тими обома вимірюваннями, які потрібно порівняти, інакше воно не дасть достатньо повної сполучної інформації.

На практиці використовують два різні підходи до якірного вимірювання. Перший підхід передбачає, що якірне вимірювання має спільний для обох груп розподіл. Це фактично означає, що обидві групи належать до однієї гіпотетичної популяції. Цей підхід використовується, наприклад, для порівняння результатів предметних тестів SAT II. Різні особи складають тести з різних предметів, а якірним є спільний для всіх тест SAT I. Другий підхід полягає у тому, що результати тієї групи екзаменованих, яким пред'являвся тест Y і якірне вимірювання A, використовуються для визначення *функції порівняння*, такої, як еквіпроцентильна функція чи лінійна

функція порівняння, для порівняння тесту *X* та вимірювання *A*. Таким чином, результати обох тестів *X* та *Y* трансформуються у шкалу якірного вимірювання *A*.

Пряме прирівнювання. Метою прямого прирівнювання є інтерпретація результатів різних тестів у такий спосіб, ніби вони є результатами одного і того ж вимірювання. Для цього необхідно, щоб обидва тести вимірювали один і той же конструкт на одному і тому ж рівні трудності і з однаковою надійністю. Це найбільш строга форма порівняння результатів двох вимірювань. Пряме прирівнювання є необхідною частиною будь-якої програми тестування, яка передбачає багатократне тестування за допомогою різних форм тестів. Хоча різні форми тесту можуть базуватися на одних і тих же специфікаціях тестових завдань і тестів, вони все ж зазвичай мають різні статистичні характеристики, наприклад, одна форма може бути в цілому складніша за іншу. головна мета прямого прирівнювання – уникнути ефектів, які є наслідками розбіжностей в статистичних характеристиках форм.

Збір даних для порівняння тестів. Існує багато методів (дизайнів) збору даних для порівняння тестів. Розглянемо основні чотири дизайни, подавши їх у вигляді таблиць.

1. Дизайн *єдиної групи* екзаменованих (SG, Single Group). Таблиця 11.1 демонструє схему цього дизайну. Тут і в наступних таблицях *X* , *Y* і *A* означають тести, *P* і *Q* – популяції. У цьому дизайні одна й та ж група екзаменованих проходить обидва тести.

Таблиця 11.1. SG-дизайн

Популяція	Вибірка	*X*	*Y*
P	1	+	+

2. Дизайн *еквівалентних груп* (EG, таблиця 11.2). Тут є дві вибірки з однієї популяції, кожна з груп проходить свій відмінний тест. Якірною інформацією є належність обох груп до однієї вибірки.

Таблиця 11.2. EG-дизайн

Популяція	Вибірка	*X*	*Y*
P	1	+	
P	2		+

3. Дизайн *рівноваги* (CB, Counterbalanced). Таблицею 11.3 представлено дизайн з двома групами із однієї популяції.

Таблиця 11.3. CB-дизайн

Популяція	Вибірка	X_1	Y_1	X_2	Y_2
P	1	+			+
P	2		+	+	

Цей дизайн використовується для того, щоб усунути ефект послідовності у пред'явленні тестів у дизайні SG. Якщо тести пред'являти одній і тій же групі у різній послідовності, результати можуть істотно відрізнятися. У дизайні CB група розбивається на дві підгрупи. Одна підгрупа виконує спочатку тест *X*, потім тест *Y* (позначається як X_1, Y_2), інша – навпаки (X_2, Y_1). Як бачимо, фактично цей дизайн є поєднанням двох попередніх – SG і EG.

4. Дизайн якірного тесту (*A*) для нееквівалентних груп (NEAT, таблиця 11.4).

Таблиця 11.4. NEAT-дизайн

Популяція	Вибірка	*X*	*A*	*Y*
P	1	+	+	
Q	2		+	+

Якірний тест може входити як частина в обидва тести *X* і *Y*, а може й бути окремим тестом. Якщо $P = Q$, то отримаємо ще один дизайн, який називають EG-дизайном з якірним тестом. У цьому випадку тест А може вимірювати не обов'язково той же конструкт, що й *X* та *Y*, але повинен тоді корелювати з ними.

ЗРАЗКИ ЗАВДАНЬ ПІДСУМКОВОГО ТЕСТУ

1. Вкажіть відповідність між величинами і шкалами, у яких ці величини виміряні:

Величини	Шкали
Номер телефону	Відношень
Сортність пшениці	Порядкова
Температура за Фаренгейтом	Категоріальна
Температура за Кельвіном	Інтервальна

2. Для кожного з перетворень виберіть шкалу, для якого це перетворення є допустимим:

Допустиме перетворення	Шкала
Перейменування значень шкали	Відношень
Заміна кожного значення шкали його десятковим логарифмом	Порядкова
	Категоріальна
Додавання до кожного значення шкали числа 10	Інтервальна
Множення кожного значення шкали на 10	

3. Вкажіть відповідність між типами шкал вимірювання і рівнями наукового усвідомлення:

Тип шкали	Рівень усвідомлення
Порядкова	Рівень вимірювання ознаки об’єкта
Категоріальна	Рівень порівняння понять
Інтервальна	Рівень утворення понять

4. До якої групи належить кожна з вибіркових характеристик вимірюваної ознаки?

Характеристика	Група
Медіана	Міри положення
Стандартне відхилення	Міри мінливості
Ексцес	Міри форми розподілу

5. Якщо коефіцієнт кореляції Пірсона дорівнює 0,9, це свідчить, що між змінними існує лінійний статистичний зв'язок, який є...

a) сильним і прямо пропорційним
b) слабким і прямо пропорційним
c) сильним і обернено пропорційним
d) слабким і обернено пропорційним

6. Значення вибіркової оцінки ексцесу 0,5 свідчить про те, що розподіл ознаки у вибірці є...

a) більш гостовершинним, ніж при нормальному розподілі
b) менш гостовершинним, ніж при нормальному розподілі
c) зсунутість вершини розподілу вправо
d) зсунутість вершини розподілу вліво

7. У якому з наступних тестів репрезентативна вибірка об'ємом у 30 осіб найбільш повно представляє популяцію?

a) Тест на уміння перемножувати двозначні числа.
b) Тест на уміння множити числа.
c) Тест на уміння виконувати арифметичні дії.
d) Тест навчальних досягнень з математики за програмою середньої школи.

8. Нехай випадкова величина розподілена нормально з математичним очікуванням 10 і стандартним відхиленням 2. Тоді близько 95% усіх випадків реалізації цієї випадкової величини потрапляє в інтервал...

a) [6, 14]
b) [0, 8]
c) [10, 18]
d) [2, 10]

9. Яку з наступних величин слід вважати виміряною в шкалі інтервалів?

a) Рік народження учня
b) Вага учня
c) Кількість учнів у кімнаті
d) Оцінка учня за активність на уроці

10. Які з наступних величин слід вважати виміряними в порядковій шкалі?

a) Місце, зайняте спортсменом на змаганнях.
b) Сума оцінок суддів для спортсмена у змаганні з фігурного катання на ковзанах.
c) Група крові людини
d) Колір очей першого за списком учня класу

11. Чим вимірювання в психології та інших суспільних науках відрізняється від вимірювання у побутовому розумінні?

a) Немає чіткої границі між суб'єктом вимірювання та інструментом вимірювання.
b) В суспільних науках розглядаються «нефізичні» величини, які не можна виміряти.
c) В суспільних науках міра якоїсь величини не є адитивною.
d) У суспільних науках вимірювання здійснюється значно простіше – «на око».

12. Чим є визначення правил відповідності між теоретичним конструктом і множиною типів поведінки?

a) Операційним визначенням конструкту
b) Тестом
c) Педагогічним рішенням
d) Шкалою вимірювання

13. Яку з наступних величин можна вважати виміряною у категоріальній шкалі?

a) Номери на футболках гравців футбольної команди
b) Температура повітря у кімнаті
c) Відстань між двома містами
d) Номери будинків вздовж вулиці

14. Які з цих множин утворюють шкалу?

a) Математична система з відношеннями
b) Емпірична система з відношеннями
c) Група допустимих перетворень шкали
d) Множина гомоморфних відображень емпіричної системи в математичну

15. У яких з цих видів тестування перевіряються оптимальні дії тестованих?

a) Тест навчальних досягнень з хімії
b) Тест на вербальні здібності
c) Тест на рівень тривожності
d) Тест на схильність до лідерства
e) Тест на комунікативність

16. У яких з цих видів тестування перевіряється типова поведінка тестованих?

a) Тест навчальних досягнень з хімії
b) Тест на вербальні здібності
c) Тест на рівень тривожності
d) Тест на схильність до лідерства
e) Тест на комунікативність

17. Яка з теорій вимірювання менше уваги приділяє властивостям об'єктів, що вимірюються, а більше дбає про логічність дій, які репрезентують ці властивості у вигляді чисел?

a) Операціональна
b) Репрезентативна
c) Класична
d) Метрологічна

18. Яке з наступних тверджень є НЕправильним?

a) Для певного конструкту існує лише одне операційне визначення
b) Психологічні вимірювання завжди базуються на обмежених вибірках спостережень
c) Вимірювання психологічних конструктів завжди супроводжуються помилками
d) Не існує природного нуля та одиниці вимірювання на обраній шкалі
e) Психологічний конструкт не може визначатися лише в термінах операційного визначення, він також повинен проявлятися у зв’язках з іншими конструктами

19. Чому дорівнює відносна частота оцінки 7 балів за певний тест, якщо цю оцінку отримали 13 учнів у вибірці з 50 учнів? Відповідь уведіть у вигляді десяткового дробу.

20.Знайдіть дисперсію дискретної випадкової величини з таким розподілом:

X	-1	0	1
P	0,2	0,6	0,2

Відповідь уведіть у вигляді десяткового дробу з одним знаком після коми.

21. Яка характеристика розподілу ознаки в популяції позначається англійським терміном variance?

22. Вибірка осіб, яка без спотворень представляє рівень вимірюваного конструкту в популяції, називається ... вибіркою.

23. Вкажіть відповідність між етапами побудови інтерпретаційного аргументу і їх назвами:

Зміст етапів	Назви етапів
Від спостережених відповідей – до тестових балів	Екстраполяція
Від тестових балів – до балів за популяцію тестових завдань	Скоринг
Від балів за популяцію тестових завдань – до балів за цільову область вимірювання	Генералізація
Від балів за цільову область вимірювання – до словесного описання рівнів вираженості риси чи конструкту	Імплікація

24. Вкажіть відповідність між методами дослідження та видами валідності:

Метод дослідження	Вид валідності
Порівняння результатів тестування з результатами критеріальних вимірювань	Дискримінативна
Порівняння результатів тестування з іншими видами оцінювання того ж самого конструкту	Конкурентна, прогностична
Порівняння результатів тестування з оцінками дивергентних конструктів	Конвергентна

25. Вкажіть відповідність між припущеннями та етапами побудови інтерпретаційного аргументу при дослідженні валідносі вимірювання рис:

Припущення	Етап
Результати тестування відповідають обраній моделі шкалування Завдання тесту є репрезентативною вибіркою з популяції тестових завдань Відсутні систематичні похибки вимірювання Вербальна інтерпретація тестових балів відповідає вимірюваній рисі	Екстраполяція Скоринг Генералізація Імплікація

26. Вкажіть логічний порядок побудови інтерпретаційного аргументу для дослідження валідності вимірювання рис:

a) Скоринг
b) Генералізація
c) Екстраполяція
d) Імплікація

27. Який з видів валідності тесту досліджується швидше на основі суджень експертів, ніж на кореляційному аналізі?

a) Змістова
b) Конкурентна
c) Прогностична
d) Конвергентна

28. Метод побудови кореляційної матриці «багато рис - багато методів» (Multitrait-Multimethod) використовується для дослідження ... валідності тесту.

a) дискримінантної
b) прогностичної
c) конкурентної
d) змістової

29. Висока кореляція між балами ЗНО з математики випускників школи та середніми балами за 1 курс їх майбутнього навчання в університеті свідчить про ... валідність тесту.

a) прогностичну
b) конкурентну

c) конвергентну
d) змістову

30. Відповідність передбачуваної інтерпретації та використання результатів тестування тій меті, заради якої створено тест - це ... тесту.

31. Співставлення змісту тестових завдань із змістом цільової області вимірювання виконується при дослідженні ... валідності тесту.

32. Вкажіть відповідність між видами надійності і джерелами похибок вимірювання:

Вид надійності	Джерело похибок
Ретестова надійність	Вибірка часових інтервалів
Надійність паралельних форм тесту (безпосередня)	Вибірка часових інтервалів плюс неоднорідність змісту
Альфа Кронбаха та надійність за К'юдером-Річардсоном	Вибірка змісту плюс неоднорідність змісту
Надійність паралельних форм тесту, пред'явлених з інтервалом у часі	Вибірка змісту

33. Вкажіть відповідність між змістом формул для обчислення коефіцієнта надійності та їх назвами:

Зміст формули	Назва коефіцієнта надійності
Компенсація впливу зменшення кількості завдань при застосуванні методу розщеплення тесту на дві еквівалентні половини	Спірмена-Брауна
Врахування усіх можливих способів розщеплення для тесту з завданнями, що оцінюються за дихотомічною шкалою	К’юдера-Річардсона
Врахування усіх можливих способів розщеплення для тесту з завданнями, для яких передбачаються частково правильні відповіді	Альфа Кронбаха

34. Для кожного методу вкажіть відповідний тип надійності тесту.

Метод	Тип надійності
Тест пред'являється двічі одній і тій же групі осіб Групі осіб по черзі пред'являються дві альтернативні форми тесту Тест розщеплюють на дві схожі за статистичними характеристиками частини, потім досліджують кореляцію між ними	Ретестова надійність Надійність розщеплених форм Надійність еквівалентних половин

35. Якщо розщіплювати тест на дві половини усіма можливими способами і обчислювати щоразу коефіцієнт кореляції між їх результатами, то середнє значення отриманих чисел буде таким же, як при використанні формули...

a) KR-20
b) Бернуллі
c) Спірмена-Брауна
d) Пуассона

36. Для яких методів обчислення надійності достатньо існування однієї форми тесту?

a) Ретестовий
b) Розщеплення на еквівалентні половини
c) KR-20 та альфа Кронбаха
d) Паралельних форм

37. Чим більшою є вибірка екзаменованих, тим ... є значення коефіцієнта кореляції, отриманий для двох ознак цих екзаменованих.

a) більш значущим
b) більшим за модулем
c) меншим за модулем
d) менш значущим

38. Формула $\rho_{XX'} = \frac{n}{n-1} \cdot \frac{s_X^2 - \sum_{i=1}^{n} s_i^2}{s_X^2}$ називається формулою...

a) альфа Кронбаха

b) Спірмена-Брауна
c) К'юдера-Річардсон
d) Пірсона

39. Які з методів обчислення надійності вимагають двох двох сеансів тестування?

a) Ретестовий
b) Паралельних форм з часовим інтервалом
c) Розщеплення на еквівалентні половини
d) Метод KR-20 та альфа Кронбаха

40. Якщо коефіцієнт надійності, обчислений як коефіцієнт кореляції між спостереженими оцінками за паралельні форми тесту, дорівнює 0,5, то вклад похибок вимірювання у дисперсію спостережених оцінок дорівнює ... відсотків (уведіть число).

41. У класичній моделі тестової оцінки постулюється, що кореляція між істинною і помилковою компонентами спостереженої оцінки дорівнює ... (уведіть число).

42. У класичній моделі тестової оцінки постулюється, що середнє значення похибок оцінювання для популяції дорівнює ... (уведіть число).

43. У класичній моделі тестової оцінки $X = T + E$ символом E позначається випадкова ... вимірювання.

45. Що означає величина P_l у формулі індексу дискримінативності $D = P_u - P_l$?

a) Частку тих з більш слабкої підгрупи екзаменованих, які відповіли на дане завдання правильно
b) Частку тих з більш сильної підгрупи екзаменованих, які відповіли на дане завдання правильно
c) Частку тих з більш слабкої підгрупи екзаменованих, які відповіли на дане завдання неправильно
d) Частку тих з більш сильної підгрупи екзаменованих, які відповіли на дане завдання правильно

46. Що означає величина $\bar{X}_+$ у формулі точково-бісеріального коефіцієнту $\rho_{pbis} = \frac{\bar{X}_+ - \bar{X}}{s_X}\sqrt{p/q}$?

a) Середню оцінку за тест тієї підгрупи екзаменованих, яка відповіла на дане завдання правильно
b) Середню оцінку за тест тієї підгрупи екзаменованих, яка відповіла на дане завдання неправильно
c) Середню оцінку за тест більш сильної підгрупи екзаменованих
d) Середню оцінку за тест більш слабкої підгрупи екзаменованих

47. Якщо тестове завдання має трудність 0,6, і на нього відповідали 10 осіб, то у скількох парах осіб вимірюваний рівень відрізняється? (Уведіть число)

48. Нехай тестове завдання виду MCQ має трудність 0,5 і ймовірність вгадати правильну відповідь на нього становить . Скільки екзаменованих з десяти у середньому відповідатимуть на це завдання правильно? (Уведіть число)

49. Якщо трудність тестового завдання з дихотомічною оцінкою {0, 1} дорівнює 0,5, то дисперсія відповідей на нього дорівнює... (Уведіть число у вигляді десяткового дробу)

50. Нехай у тесті 5 є завдань з дихотомічною шкалою оцінювання і трудність завдань становить, відповідно, 0,4; 0,5; 0,6; 0,7; 0,8.Чому дорівнює середня трудність завдань? (Уведіть число)

51. Нехай у тесті 5 є завдань з дихотомічною шкалою оцінювання і трудність завдань становить, відповідно, 0,4; 0,5; 0,6; 0,7; 0,8.Чому дорівнює трудність тесту? (Уведіть число)

ЛІТЕРАТУРА

1. Анастази А., Урбина С. Психологическое тестирование. – 7-е изд. – СПб.: Питер, 2007. – 688 с.: ил. – (Серия «Мастера психологии»). – ISBN 978-5-272-00106-1.

2. Булах І. Є. Створюємо якісний тест : навч. посібник / І. Є. Булах, М. Р. Мруга. – К.: Майстер-клас, 2006. – 160 с.

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

4. Звонников В.И., Челышкова М.Б. Современные средства оценивания результатов обучения. – М.: Издательский центр «Академия». 2007. – 223 с.

5. Ким В.С. Тестирование учебных достижений. Монография. – Уссурийск: Изд-во УГПИ, 2007. – 214 с.

6. Крокер Л., Алгина Дж. Введение в классическую и современную теорию тестов. – М.: Логос, 2010. – 668 с.

7. Нейман Ю.М., Хлебников В.А. Введение в теорию моделирования и параметризации педагогических тестов. – М.: Прометей, 2000. – 168 с.

8. Педагогічне оцінювання і тестування. Правила, стандарти, відповідність. Наукове видання / Я.Я.Болюбаш, І.Є.Булах, М.Р.Мруга, І.В.Філончук. – К.: Майстер-клас, 2007.– 272 с.

9. Челышкова М.Б. Теория и практика конструирования педагогических тестов. – М.: Логос, 2002. – 432 с.

10. Educational Measurement. – 4th edition / Edited by Robert L. Brennan. – ACE, 2006. – 796 pp.

11. Wright B.D., Stone M.H. Best Test Design. – Chicago, MESA PRESS, 1979. – 222 pp.