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.

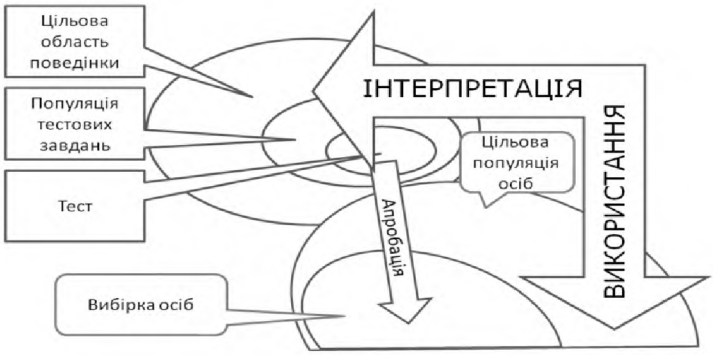

Рис. 4.1. Співвідношення множин суб'єктів та об'єктів тестування

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 $\rho_{i^*}^2$ 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.