Тема 2. Психометричні засади: надійність, валідність, стандартизація, норми

План лекції

2.1. Вимірювання в психології, надійність і її показники

2.2. Валідність, стандартизація та норми: як сирий бал набуває змісту

2.3. Похибка вимірювання, обмеження методик у особливих групах і роль скринінгу

Література: [7, 9, 14, 16, 20, 21, 22, 25]

2.1. Вимірювання в психології, надійність і її показники

Психологічне вимірювання полягає в приписуванні числових значень властивостям людини за визначеними правилами так, щоб відношення між числами відображали відношення між властивостями. Психічні конструкти (тривожність, інтелект, життєстійкість) є латентними: їх оцінюють за сукупністю поведінкових індикаторів, а не безпосередньо [14; 16]. За класифікацією С. Стівенса розрізняють чотири рівні шкал: номінальну (класифікація без порядку; допустимі частоти й мода), порядкову (ранжування без рівних інтервалів; медіана, рангові кореляції), інтервальну (рівні інтервали без абсолютного нуля; середнє, стандартне відхилення, параметричні критерії) і відношень (є абсолютний нуль). Сумарні бали опитувальників відносять до порядкових або умовно інтервальних, тому припущення про рівність інтервалів потребує перевірки, а в разі сумніву віддають перевагу непараметричним методам [22].

Теоретичну основу оцінювання якості вимірювання створює класична теорія тестів, яка виходить з рівняння:

$$X = T + E$$

де: X – спостережуваний бал; T – істинний бал; E – випадкова похибка, середнє якої дорівнює нулю й яка не корелює з T.

Звідси надійність визначається як частка дисперсії істинних балів у загальній дисперсії спостережуваних балів. Вона дорівнює квадрату кореляції між спостережуваним та істинним балом. Чим менша випадкова похибка, тим стабільніші результати. Недолік класичної моделі – залежність характеристик тесту від вибірки й однакова точність для всіх рівнів ознаки.

Альтернативою класичній моделі є теорія відповіді на завдання (IRT), яка моделює ймовірність відповіді залежно від латентної змінної та параметрів завдання. Однопараметрична модель (модель Г. Раша) враховує лише складність завдання, двопараметрична додає дискримінативність, трипараметрична – ймовірність вгадування. Інформаційна функція показує точність вимірювання на кожному рівні ознаки, що відрізняє IRT від класичної теорії з єдиним коефіцієнтом надійності. Аналіз диференційного функціонування завдань виявляє пункти, які по-різному працюють у різних групах за однакового рівня ознаки. На IRT ґрунтується комп’ютерне адаптивне тестування. Узагальнена теорія (теорія генералізованості) розкладає дисперсію результатів за джерелами – обстежувані, пункти, умови, оцінювачі, – і дає змогу встановити, який чинник найбільше знижує точність.

Практично надійність розглядають як властивість методики давати узгоджені результати за повторних або еквівалентних вимірів, а вид надійності обирають відповідно до джерела похибки. Ретестова надійність оцінює стабільність у часі як кореляцію результатів двох вимірів з інтервалом, достатнім для стирання слідів пам’яті, але недостатнім для реальних змін. Для станових шкал (тривога, симптоматика) високої ретестової надійності не очікують. Надійність паралельних форм перевіряє еквівалентність двох варіантів методики, а джерелом похибки є вибірка пунктів. Внутрішня узгодженість показує, наскільки пункти шкали вимірюють один конструкт. Надійність розщеплення – кореляція двох половин тесту, до якої застосовують поправку Спірмена–Брауна:

$$r = 2r_{½} / (1 + r_{½})$$

де: r – надійність усього тесту; $r_{½}$ – кореляція між його половинами [9; 20].

Довжина тесту впливає на надійність: що більше еквівалентних пунктів, то вища надійність. За узагальненою формулою Спірмена–Брауна можна оцінити, скільки пунктів потрібно для заданої надійності й скільки надійності втрачається в скороченій версії. Припущення формули – однорідність доданих пунктів [16; 22].

Найпоширенішим показником внутрішньої узгодженості є коефіцієнт альфа Л. Кронбаха (1951): він тим вищий, чим сильніше пов’язані між собою пункти шкали й чим їх більше. Для дихотомічних пунктів його аналогом є формула Кюдера–Річардсона (KR-20).

Значення альфа 0,70–0,80 вважають прийнятними для групових досліджень, понад 0,90 – для індивідуальних рішень із тяжкими наслідками. Значення понад 0,95 свідчить про надлишковість пунктів. Альфа зростає зі збільшенням кількості пунктів і середньої міжпунктової кореляції, тому порівнювати скорочену й повну версії за цим коефіцієнтом некоректно. Високе значення альфа не доводить одновимірності шкали, для чого потрібен факторний аналіз. Коефіцієнт альфа припускає однакові факторні навантаження пунктів, тому дедалі частіше використовують коефіцієнт омега Р. Макдональда, що спирається на фактичні навантаження.

Аналіз пунктів передує оцінюванню надійності. Індекс складності (частка правильних відповідей) і індекс дискримінативності (різниця між групами високих і низьких результатів або кореляція пункту із сумою) визначають придатність завдань тесту досягнень та тесту інтелекту. Для опитувальників обчислюють скориговану кореляцію пункту із сумою за вилученим пунктом (бажано не менше 0,30) і значення альфа в разі вилучення пункту. Пункт, вилучення якого підвищує альфа, перевіряють на зміст і формулювання. Середня міжпунктова кореляція в межах приблизно 0,15–0,50 вважається показником узгодженості без надмірної надлишковості.

Надійність різницевих балів (наприклад, різниці між двома субшкалами профілю) нижча за надійність кожної зі шкал, особливо коли шкали сильно пов’язані між собою.

За високої кореляції шкал різницевий бал практично ненадійний, тому розбіжності профілю інтерпретують лише тоді, коли вони перевищують критичні значення, наведені в посібнику методики. Значення надійності, наведене в посібнику, не переноситься автоматично на нову вибірку, особливо за іншого розкиду ознаки, тому надійність перевіряють у вибірці користувача.

Міжоцінювальна надійність стосується методів, де результат залежить від судження діагноста: структурованих інтерв’ю, спостереження, оцінки проєктивних відповідей. Для категоріальних оцінок застосовують коефіцієнт каппа Дж. Коена (згода з поправкою на випадкові збіги). За шкалою Дж. Ландіса й Г. Коха значення 0,41–0,60 означає помірну, 0,61–0,80 – суттєву, понад 0,80 – майже повну згоду. Для кількісних оцінок використовують внутрішньокласовий коефіцієнт кореляції (значення понад 0,75 розглядають як добру, понад 0,90 – як відмінну надійність). Згоду підвищують чіткі операціональні критерії, навчання оцінювачів та калібрування на спільних випадках [14; 16].

Рівні надійності відомих інструментів дають орієнтири для оцінювання нових методик. Опитувальник PHQ-9 має внутрішню узгодженість 0,86–0,89 і ретестову надійність близько 0,84 за 48 годин; шкала GAD-7 – альфу близько 0,92; чек-лист PCL-5 – близько 0,94–0,96 у вибірках осіб, що пережили травму. Для коротких субшкал із трьох-чотирьох пунктів надійність за означенням нижча, ніж для інтегральної шкали, тому її оцінюють окремо для кожної субшкали.

Чинники, що знижують надійність: неоднозначні формулювання, мала кількість пунктів, неоднорідність змісту, неточний підрахунок; шум, втома, низька мотивація, неправильне розуміння інструкції, зміни стану; обмеження діапазону ознаки у вибірці (у відібраній групі коефіцієнти кореляції та

надійності нижчі, ніж у загальній популяції). Надійність належить не лише методиці, а й конкретному застосуванню, тому стандартні умови проведення є частиною психометричної якості процедури [9; 20].

2.2. Валідність, стандартизація та норми: як сирий бал набуває змісту

Валідність визначає, наскільки результати методики обґрунтовують запропоноване тлумачення. За сучасним розумінням (С. Мессік, 1989; Стандарти педагогічного й психологічного тестування, 2014) валідність є не властивістю тесту, а характеристикою висновків, зроблених на підставі його результатів у визначеному контексті. Вона підтверджується сукупністю доказів п’яти типів: за змістом методики, процесами відповіді, внутрішньою структурою, зв’язками з іншими змінними та наслідками застосування. Методика, валідна для відбору кандидатів, може бути невалідною для оцінювання результатів психотерапії.

Крім валідності методики, перевіряють якість критерію. Критерій має бути релевантним меті, надійним і вільним від забруднення (на оцінку критерію впливає знання результатів тесту) та дефіциту (критерій не охоплює важливих аспектів діяльності). У прогностичних дослідженнях за відібраних вибірок діапазон ознаки обмежений, що занижує кореляцію методики з критерієм. Для оцінювання валідності в усій популяції кандидатів застосовують поправку на обмеження діапазону. Рівняння регресії та правила ухвалення рішень, побудовані на одній вибірці, завищують точність (ефект зменшення, або «усадки»), тому їх обов’язково перевіряють на незалежній вибірці [22].

Докази валідності мають різну природу. Змістова валідність показує, що пункти репрезентативно охоплюють область конструкту. Її оцінюють експерти, а ступінь згоди щодо необхідності кожного пункту виражають коефіцієнтом змістової валідності Ч. Лоуші. Критеріальна валідність демонструє зв’язок з іншим показником: одночасна (кореляція з чинним критерієм, зокрема діагностичним інтерв’ю) і прогностична (здатність передбачати майбутній

критерій, зокрема ефективність виконання службових (бойових, рятувальних, оперативних) завдань). Валідність за контрастними групами підтверджується, якщо методика розрізняє групи, між якими різницю передбачає теорія (особи з діагностованим розладом і без нього).

Надійність обмежує валідність: кореляція методики з критерієм не може бути вищою за корінь із добутку їх надійностей, тому низька надійність обмежує можливу валідність.

Додаткова (інкрементна) валідність виражається приростом пояснюваної дисперсії критерію (приріст R^2) від включення методики до набору наявних предикторів. Метод із низькою додатковою валідністю не виправдовує витрат на його застосування.

Конструктна валідність (Л. Кронбах, П. Мілл, 1955) охоплює збіжну й розбіжну валідність: нова шкала тривожності має суттєво корелювати з визнаною шкалою тривожності (збіжність) і слабко – із суміжними конструктами (розбіжність). Д. Кемпбелл і Д. Фіске (1959) запропонували багатоознаково-багатометодну матрицю: кореляції одного конструкту, виміряного різними методами (діагональ валідності), мають перевищувати кореляції різних конструктів, виміряних одним методом, що відокремлює змістову дисперсію від дисперсії методу [14].

Факторна валідність перевіряє внутрішню структуру. Розвідувальний факторний аналіз виявляє кількість і склад факторів, а підтверджувальний перевіряє відповідність заданої моделі даним за індексами узгодженості: CFI і TLI не менше 0,95, RMSEA не більше 0,06, SRMR не більше 0,08 (критерії Л. Ху і П. Бентлера). Такі процедури є ядром адаптації методик: психометричну валідизацію української версії шкали постійної травматичної реакції проведено з перевіркою факторної структури, надійності й зв’язків з іншими змінними в умовах війни [25], а національні вибірки батьків і дітей слугували для перевірки міжнародних інструментів діагностики ПТСР і комплексного ПТСР.

Обсяг вибірки для факторного аналізу залежить від кількості пунктів і величини навантажень: за орієнтирами не менше 200 респондентів і 5–10 респондентів на один пункт. За менших навантажень вибірка має бути більшою. Для дихотомічних і порядкових відповідей коректнішим є аналіз на основі поліхоричних кореляцій, а не звичайних коефіцієнтів Пірсона [16; 22].

Для скринінгових шкал критеріальну валідність виражають через діагностичну точність, яку визначають зіставленням з еталонним критерієм (структурованим інтерв'ю). Чутливість – частка виявлених серед осіб зі станом; специфічність – частка правильно відхилених серед осіб без стану:

$$Se = ІП / (ІП + ХН); Sp = ІН / (ІН + ХП)$$

де: Se – чутливість; Sp – специфічність; ІП – істинно позитивні результати; ХН – хибно негативні; ІН – істинно негативні; ХП – хибно позитивні.

Прогностична цінність позитивного (PPV) і негативного (NPV) результату – це частка справжніх випадків серед осіб із позитивним результатом і частка осіб без стану серед осіб із негативним результатом. Прогностична цінність залежить від поширеності стану. Відношення правдоподібності показує, у скільки разів позитивний результат частіше трапляється в осіб зі станом, ніж без нього.

Щоб оцінити точність за всіма можливими порогами відсічення, будують ROC-криву. Площа під нею (AUC) показує загальну розрізнювальну здатність: 0,5 – рівень випадку, 0,7–0,8 – прийнятна, 0,8–0,9 – добра, понад 0,9 – відмінна. Оптимальний поріг обирають за балансом чутливості й специфічності та за клінічною логікою.

Формальним критерієм балансу є індекс Юдена (J = чутливість + специфічність – 1): оптимальним вважають поріг з його найбільшим значенням. Поріг PHQ-9 у 10 балів забезпечує чутливість і специфічність близько 0,88

щодо великого депресивного епізоду. Для PCL-5 у різних вибірках пропонують порогові значення 31–33 бали залежно від популяції та мети [14].

Точність, отримана в клінічній вибірці з виразною симптоматикою, зазвичай вища за точність у загальній популяції, де випадки легші і менш чітко відмежовані від норми (зміщення спектра). Звітування про дослідження діагностичної точності здійснюють за настановою STARD, яка вимагає опису вибірки, еталонного критерію, порогів, чутливості й специфічності з довірчими інтервалами. Для критеріально-орієнтованих методик поріг може встановлюватися експертним методом Ангофа: експерти оцінюють для кожного завдання ймовірність правильної відповіді мінімально компетентного екзаменованого, а середня сума цих оцінок задає поріг.

Загрози валідності: недостатня репрезентативність змісту, наявність у пунктах стороннього змісту, соціальна бажаність, культурна упередженість, мовні відмінності, вплив ситуаційних чинників, невідповідність мети застосування меті створення методики. Валідність, доведена на цивільній вибірці, не переноситься автоматично на військових, рятувальників чи ветеранів, а доведена для скринінгу – на експертну діагностику [9; 20].

Два інші типи доказів стосуються процесів відповіді та наслідків застосування. Докази за процесами відповіді отримують когнітивним інтерв’юванням («мислення вголос»): респондентів просять пояснити, як вони розуміють твердження й обирають відповідь. Розбіжність між задумом автора й розумінням респондентів свідчить про необхідність перегляду формулювання. Докази за наслідками застосування стосуються того, чи досягає використання результатів мети і чи не спричиняє непередбачуваних негативних наслідків, зокрема систематичного відсіву певної групи кандидатів.

Щоб сирий бал набув змісту, потрібні стандартизація та норми. Стандартизація – уніфікація умов проведення, інструкції, часових обмежень, матеріалів, правил підрахунку й інтерпретації. Вона забезпечує порівнянність результатів різних осіб і різних обстежень. Порушення стандартизації (зміна

інструкції, допомога під час виконання, зміна часових обмежень) знижує валідність нормативної інтерпретації, і про такі відхилення зазначають у висновку.

Норми – розподіл балів у нормативній вибірці. Вибірка має бути репрезентативною за віком, статтю, освітою, культурою та умовами діяльності й достатньою за обсягом у кожній віковій чи соціальній страті. Нормовані показники: перцентиль – частка вибірки з нижчим балом. Решту показників обчислюють перетворенням сирого бала на z-оцінку:

$$z = (X - M) / SD$$

$$T = 50 + 10z;\ КІ = 100 + 15z;\ стен = 5{,}5 + 2z;\ станайн = 5 + 2z$$

де: z – стандартна z-оцінка (середнє 0, відхилення 1); X – сирий бал; M – середнє нормативної вибірки; SD – стандартне відхилення нормативної вибірки; T – T-бал (середнє 50, відхилення 10); КІ – відхилений коефіцієнт інтелекту (середнє 100, відхилення 15; шкали Д. Векслера); стен – десятибальна шкала (середнє 5,5, відхилення 2); станайн – дев'ятибальна шкала (середнє 5, відхилення 2).

За нормального розподілу $z = \pm 1$ відповідає приблизно 84-му й 16-му перцентилям, $z = \pm 2$ – приблизно 98-му й 2-му [22].

Перетворення на стандартні шкали передбачає нормальний розподіл. Для скошених розподілів застосовують нормалізовані оцінки. Перцентилі нерівномірно відображають відмінності: у центрі розподілу невелика різниця балів дає значну різницю перцентилів, на краях – незначну. Застарілі норми зумовлюють систематичне зміщення результатів. Для когнітивних тестів описано ефект Флінна – зростання середніх результатів приблизно на 3 бали коефіцієнта інтелекту за десятиріччя, що вимагає періодичного перегляду норм.

Обсяг нормативної вибірки за практичними орієнтирами становить кілька сотень осіб на страту, а для вікових норм дитячих і когнітивних тестів нерідко застосовують безперервне нормування (регресійне згладжування залежності

бала від віку). Нормативну вибірку описують у посібнику за обсягом, складом, регіонами й роком збору. Відсутність цих відомостей унеможливлює оцінку придатності норм для конкретної групи користувача.

Критеріальне оцінювання відрізняється від нормативного: бал порівнюють не з групою, а із заданим стандартом (поріг клінічної значущості, мінімальний рівень придатності). Для скринінгових інструментів у професійних групах (військових, рятувальників, поліцейських, медиків) потрібні як загальнопопуляційні, так і специфічні норми, отримані на вибірках відповідної групи. Перенесення цивільних норм без перевірки може занижувати або завищувати частку осіб із симптоматикою [7].

На придатність шкали впливають також підлоговий і стельовий ефекти: вони виникають, коли значна частка обстежуваних (понад 15 %) набирає мінімальний чи максимальний бал: шкала втрачає чутливість до відмінностей у цій частині діапазону. Для скринінгових шкал у загальній популяції підлоговий ефект очікуваний і не є вадою, а для моніторингу в клінічних групах – обмеження [16].

Окремої уваги потребує упередженість методики, тобто систематична різниця балів у груп за однакового рівня конструкту, не пов’язана з самим конструктом. Її виявляють аналізом диференційного функціонування пунктів та перевіркою диференційної прогностичності (різні регресійні прямі, що зв’язують бал методики з критерієм, у різних груп). Справедливість тестування вимагає однакових умов, доступності інструкцій і відсутності бар’єрів, не пов’язаних із вимірюваним конструктом.

Адаптація методики. Перенесення зарубіжної шкали в іншу культуру є окремою психометричною процедурою. Під час адаптації короткої форми шкали симптомів моральної травми (MISS-M-SF) виконали три прямі й два зворотні переклади та когнітивне пілотне опитування п’яти респондентів, після чого перевірили властивості шкали на вибірці з 111 осіб: α Кронбаха становила 0,70, ретестова кореляція (інтервал 8 днів, n = 32) – 0,67, кореляції пунктів із

загальним балом – 0,45–0,71. Дискримінантну валідність підтверджено практично нульовими кореляціями з PCL-5 (r = 0,04) і GAD-7 (r = 0,07), конвергентну – r = 0,28.

Українську адаптацію GAD-7 також перевіряли на двох вибірках цивільного населення віком 17–75 років (786 осіб у грудні 2023 року та 315 осіб у січні 2024 року) з прямим і зворотним перекладом: внутрішня узгодженість була високою (α = 0,90; ω = 0,89; спліт-половинна надійність 0,87), ретестова надійність за чотири тижні становила 0,72, а один фактор пояснював 57,1 % дисперсії [21].

Норми та надійність нової української методики. Опитувальник грамотності щодо психічного здоров’я (MHLq) містить 29 пунктів з п’ятибальною шкалою Лайкерта та чотири шкали (знання – 13 пунктів, стереотипи – 6, звернення по допомогу й перша допомога – 6, самодопомога – 4). Бали за стереотипами враховують з від’ємним знаком. На вибірці з 430 осіб віком 18–60 років чотири фактори пояснили 60,43 % дисперсії, дельта Фергюсона становила 0,91–0,95, ретестова надійність – 0,86–0,92 (загальний бал – 0,90), спліт-половинна – 0,93. Для методики розроблено процентильні норми.

2.3. Похибка вимірювання, обмеження методик у особливих групах і роль скринінгу

Стандартна похибка вимірювання обчислюється за формулою:

$$SEM = SD \times \sqrt{(1 - rxx)}$$

де: SEM – стандартна похибка вимірювання; SD – стандартне відхилення нормативної вибірки; rxx – коефіцієнт надійності.

Навколо спостережуваного бала будують довірчий інтервал: з імовірністю 95 % істинний бал лежить у межах приблизно ±1,96 SEM. Два бали, що відрізняються менше ніж на значення похибки, не можна вважати різними. Результат подають діапазоном, а не точковим значенням. Із підвищенням

надійності SEM зменшується: за rxx = 0,90 вона становить приблизно 0,32 SD, за rxx = 0,70 – приблизно 0,55 SD.

Для коротких шкал із невисокою надійністю похибка вимірювання значна, тому крайні результати коротких шкал треба тлумачити обережно.

Для оцінювання змін між повторними вимірами застосовують індекс надійної зміни Н. Джейкобсона і П. Траукса:

$$RCI = (X2 – X1) / Sдиф; Sдиф = \sqrt{2} \times SEM$$

де: RCI – індекс надійної зміни; X1 і X2 – результати першого та повторного вимірювання; Sдиф – стандартна похибка різниці; SEM – стандартна похибка вимірювання.

Значення RCI понад 1,96 вказує на зміну, що перевищує похибку вимірювання. Клінічну значущість визначають за одним із критеріїв: результат віддалився від клінічної групи більш ніж на два стандартні відхилення; потрапив у межі двох стандартних відхилень від середнього нормативної групи; перетнув умовну точку між розподілами клінічної та нормативної груп. Поєднання RCI і порога клінічної значущості відрізняє статистично достовірну зміну від клінічно значущого одужання [14].

Величину групових змін виражають розміром ефекту d – різницею середніх значень, поділеною на спільне стандартне відхилення.

Орієнтовні межі за Дж. Коеном: 0,2 – малий, 0,5 – середній, 0,8 – великий ефект. Статистична значущість залежить від обсягу вибірки, а розмір ефекту – ні, тому в оцінюванні ефективності допомоги наводять обидва показники разом з довірчими інтервалами.

Інтерпретуючи повторні виміри, враховують регресію до середнього (крайні первинні результати за повторного виміру зміщуються до середнього без втручання) та ефект практики (покращення виконання завдань тестів на швидкість і пам’ять внаслідок повторного пред’явлення). Для запобігання

цьому застосовують паралельні форми, контрольні групи й інтервали між вимірами, достатні для зниження впливу пам’яті [16].

Систематичні похибки зумовлені стилями відповідей і соціально бажаним відповіданням. Схильність до згоди з твердженнями, тенденція до крайніх або середніх відповідей, симуляція та дисимуляція спотворюють результати опитувальників. Контроль здійснюють шкалами достовірності (брехні, непослідовності, нечастих відповідей), збалансованими за напрямом формулювання шкалами, умовами проведення (анонімність, нейтральна мотивація) та зіставленням з даними інших джерел [9; 20].

Вимірювальна інваріантність показує, чи вимірює методика один і той самий конструкт у різних групах (чоловіки й жінки, військові й цивільні, різні мови). Статистична перевірка інваріантності з’ясовує, чи однакова структура методики в різних групах, адже без цього порівняння середніх значень між групами безпідставне [25].

Адаптація зарубіжної методики охоплює переклад, експертне узгодження, зворотний переклад, когнітивне інтерв’ювання, пілотування, перевірку надійності, факторної структури та валідності й побудову локальних норм. Лише переклад пунктів без психометричної перевірки не є адаптацією, а повноцінна адаптація передбачає всі перелічені етапи.

Обмеження методик у особливих групах пов’язані з віком, освітою, мовною компетентністю, станом здоров’я, культурою. Методики, нормовані на дорослих, непридатні для підлітків без відповідних норм. Результати когнітивних тестів залежать від освіти й мови. Симптоматика гострого стресу змінює показники уваги й пам’яті. У воєнних і кризових умовах вибірки неоднорідні за досвідом, а симптоми динамічні, тому висновок враховує час, що минув після події.

Скринінг відрізняється від діагностики метою й вимогами до якості. Скринінгова методика забезпечує швидке виявлення осіб, які потребують поглибленої оцінки. Критичною є висока чутливість за прийнятної

специфічності. Результат скринінгу не є діагнозом: позитивний результат вимагає підтвердження структурованим інтерв'ю або клінічною оцінкою. Прогностична цінність позитивного результату залежить від поширеності стану.

За низької поширеності навіть якісний скринінг дає значну частку хибнопозитивних результатів, і прогностична цінність позитивного результату знижується.

Практичне рішення щодо порога має два варіанти. Один поріг ділить результати на дві категорії й втрачає інформацію про осіб поблизу межі. Два пороги створюють нижню межу виключення (висока чутливість: нижче неї стан малоймовірний) і верхню межу підтвердження (висока специфічність: вище неї стан імовірний), а проміжна зона потребує подальшого обстеження. Така схема зменшує частку помилкових рішень у суперечливих випадках [14; 16].

Придатність методики оцінюють за її паспортом, який містить: назву, автора, рік і версію; конструкт і призначення; цільову вибірку; кількість пунктів і формат відповіді; правила підрахунку; нормативні дані (обсяг, склад, рік); надійність (альфа, ретест, міжоцінювальна згода) з довірчими інтервалами; докази валідності; пороги відсічення, чутливість і специфічність; шкали достовірності; обмеження застосування; вимоги до користувача; етичні застереження. Відсутність будь-якого розділу знижує довіру до результатів, особливо за використання методики для рішень щодо особи [9; 22].

Мінімальні орієнтири якості: надійність понад 0,90 для рішень із тяжкими наслідками (відбір, експертиза), понад 0,80 для скринінгу й моніторингу, понад 0,70 для групових досліджень; норми, отримані на вибірці, зіставній із досліджуваною; підтверджена критеріальна валідність за зовнішнім показником; контроль достовірності відповідей. Якщо методика не відповідає цим вимогам, її використовують як допоміжну й не розглядають результат як єдину підставу висновку.

Обмеження адаптованих методик. Українська версія опитувальника результативності психотерапії OQ-45.2 перевірена на нормативній групі з 514 осіб і ризикових групах із 114 осіб: α Кронбаха для шкал симптомного дистресу, міжособистісних стосунків, соціальних ролей і загального балу становила 0,92; 0,79; 0,71 і 0,94, ретестова надійність (N = 30, три тижні) – від 0,79 до 0,84. Водночас автори адаптації зазначають, що критеріальну еквівалентність не підтверджено (середні значення були суттєво вищими: 63 ± 27 проти 42 ± 17), факторний аналіз і перевірку чутливості до змін не проводили, клінічної вибірки не було. Цей приклад показує, що навіть добра надійність не замінює повного набору доказів валідності, а висока вираженість симптомів у воєнний час потребує обережного використання норм: так, високі бали за GAD-7 можуть відображати ситуативну тривогу, а не розлад [21].

Ключові поняття теми: *психологічне вимірювання; надійність (ретестова, внутрішня узгодженість, міжоцінювальна); коефіцієнт альфа Кронбаха; валідність (змістова, критеріальна, конструктна); факторний аналіз; стандартизація; норма, перцентиль, стандартний бал; стандартна похибка вимірювання; індекс надійної зміни; чутливість і специфічність; адаптація методики; прямий і зворотний переклад; процентильні норми.*

Питання для самоперевірки

1. Яке рівняння лежить в основі класичної теорії тестів і як з нього визначають надійність?

2. Які види надійності розрізняють і за яких умов застосовують кожен із них?

3. Чим сучасне розуміння валідності відрізняється від традиційного поділу на види?

4. Як за ROC-аналізом обирають поріг відсічення для скринінгової шкали?

5. Які види нормованих показників використовують у психодіагностиці та як вони пов’язані між собою?

6. Як за стандартною похибкою вимірювання й індексом надійної зміни тлумачать різницю між двома балами?

7. Як поширеність стану впливає на прогностичну цінність позитивного результату скринінгу та чому для суперечливих випадків використовують два пороги відсічення?

8. Які етапи адаптації зарубіжної методики та які показники підтверджують її надійність і валідність для української вибірки?