

УДК 303.214.3

ПОБУДОВА ІНТЕРПРЕТАЦІЙНОГО АРГУМЕНТУ Й АРГУМЕНТУ ВАЛІДИЗАЦІЇ У КОМПЛЕКСНОМУ ДОСЛІДЖЕННІ ВАЛІДНОСТІ СТАНДАРТИЗОВАНОГО ПРЕДМЕТНОГО ТЕСТУ

Ковальчук Ю.О.

Обґрунтована необхідність комплексної валідизації стандартизованих предметних тестів. Показана залежність змісту інтерпретаційного аргументу від мети тестування. Наведено варіанти побудови інтерпретаційного аргументу та аргументу валідності предметних тестів для випадків вимірювання риси й теоретичного конструкту.

Ключові слова: валідність вимірювання, предметний тест, інтерпретаційний аргумент, аргумент валідності.

Обоснована необходимость комплексной валидации стандартизированных предметных тестов. Показана зависимость содержания интерпретационного аргумента от цели тестирования. Приведены варианты построения интерпретационного аргумента и довода валидности предметных тестов для случаев измерения черты и теоретического конструкта.

Ключевые слова: валидность измерения, предметный тест, интерпретационный аргумент, аргумент валидности.

The necessity of complex validation of standardized subject tests is substantiated. Dependence of the interpretative argument content on the purpose of testing is established. Variants of interpretative argument and validity argument of subject tests for cases of measuring traits and theoretical constructs are set forth.

Key words: validity of measurement, subject test, interpretative argument, validity argument.

В [1] розглядався підхід М.Кейна до проблеми загальної валідизації вимірювань в освіті як до міні-теорії, що базується на побудові інтерпретаційного аргументу – системи припущень та тверджень, правильність яких необхідно обґрунтувати для того, щоб вважати вимірювання валідним. У [2] показано, що подібні теорії будуються як системи практичної аргументації і тому не можуть бути повними.

На сьогодні в Україні у вигляді “промислових”, частково стандартизованих тестів існують лише предметні тести ЗНО. Тому важливо розуміти, як повинна досліджуватися валідність, у першу чергу, предметних тестів. Таке розуміння буде сприяти, з одного боку, створенню більш якісних інструментів вимірювання, з іншого боку – сприйняттю цих інструментів стейкхолдерами – зацікавленими особами. Далі розглянемо особливості побудови інтерпретаційного аргументу для предметного тесту залежно від обраної мети тестування, а також дамо короткий огляд методів, які можуть використовуватися під час побудови відповідного аргументу валідності.

Необхідність комплексної валідизації вимірювання. Укладачі та користувачі тестів часто

обмежуються лише перевіркою часткових свідчень валідності. У деяких випадках це цілком виправданий підхід. Наприклад, якщо вчитель у 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 аспекти валідності.

Для ясності 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 окрім 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 градування 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 неправильно 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°C? Яка температура вважається нормальною, і чи є ця норма застосованою 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 подіє 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 питань, 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 валідації навіть такого простого, порівняно 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 педагогічні, 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.

Дослідження валідності як система практичної аргументації. 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 валідацію як міні-теорію, засновану на системі практичної аргументації. Розглянемо загальні принципи побудови такої теорії. За С.Тулміним [3], 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 посиляються, 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 островах, 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 упевненості 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 упевненості, ніж твердження “Я, імовірно, є громадянином Великобританії”.

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

Інтерпретаційний аргумент. Для комплексного дослідження валідності вимірювання як процесу побудови системи практичної аргументації ми повинні спочатку виробити своєрідний “план” у формі мережі тверджень, для яких потім шукатимемо необхідні свідчення. Наявність такого “плану” дозволить нам, по-перше, контролювати логіку наших дій, а по-друге, вбереже нас від ризику пропустити якийсь важливий аспект. Цей “план”, услід за М.Кейном, назовемо *інтерпретаційним аргументом*. Ті ж свідчення валідності, які будуть підкріплювати відповідні твердження інтерпретаційного аргументу, складатимуть *аргумент валідності*. Як зазначалося, особливістю освітніх вимірювань є те, що зазвичай ми хочемо інтерпретувати результати вимірювання ширше, ніж випливає безпосередньо з тестових балів. Так, у підсумковому тестуванні учнів з деякого предмету нас цікавить не “рівень успішності учня з даного предметного тесту”, а рівень його успішності з предмету взагалі, або й навіть більше – наскільки успішним буде подальше навчання учня з цього предмету, чи наскільки цей учень здатний до навчання в цілому. Укладач тесту повинен добре розуміти з самого початку, з якою метою створюється тест, адже від цього залежить зміст тієї області поведінки, яку цей тест повинен репрезентувати. Тут можна виділити два принципово відмінні випадки: або тест вимірює певну рису особи (*trait*), якою може бути, наприклад, рівень успішності з даного предмету, або – більш складний випадок – вимірюється певний теоретичний конструкт, яким є, наприклад, здібність особи до навчання взагалі, чи навіть рівень її інтелекту. У другому випадку результат тестування з окремого предмету є лише одним із можливих індикаторів конструкту. Так, загальна здібність до навчання (чи загальна навчальна компетенція) складається як мінімум з двох частин – здібності особи до логіко-математичного мислення та її здібності до вербального мислення, тому для визначення рівня цього конструкту повинні використовуватися результати тестування як з математики, так і з мови.

У будь-якому випадку, інтерпретаційний аргумент будується у тому порядку, у якому відбувається перенесення інтерпретації тестових балів з більш вузького, часткового рівня, на більш високі рівні узагальнення. Таким чином, цей порядок є оберненим до порядку укладання тесту як інструменту вимірювання. Кожне з послідовних узагальнень таїть у собі небезпеку втрати валідності вимірювання. Зауважимо, що такий порядок побудови інтерпретаційного аргументу є логічним, а не хронологічним, адже укладач тесту повинен з самого початку контролювати деякі аспекти валідності. Розглянемо детальніше структуру інтерпретаційного аргументу для випадків

вимірювання риси та теоретичного конструкту, паралельно вказуючи на методи побудови відповідних аргументів валідності.

Інтерпретаційний аргумент для вимірювання риси. Нехай потрібно дослідити валідність предметного тесту для випадку, коли метою вимірювання є з’ясування рівня успішності (в іншій термінології – рівня досягнень) особи з даного предмету. Порядок узагальнень інтерпретації тестових балів у цьому випадку є таким (у 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. Нарухування тестових балів було неупередженим.

4. Наруховані бали узгоджуються з обраною моделлю шкалування.

Аргумент валідації 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 єдиною 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 під час його переходу до іншого університету [4]. З точки зору забезпечення валідності вимірювання, шкалювання не повинне суперечити тій кінцевій меті, заради якої проводиться тестування. Так, під час застосування методу еквіпроцентильної нормалізації у ЗНО важливим є відповідність його головному принципу – той випускник, який отримав більше “сирих” балів порівняно з іншим випускником, також повинен отримати більше балів і після проведення процедури шкалювання.

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.

У разі просування 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, у разі тестування 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 (*конвергентна 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 смисловою 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 спростувати 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.

Література

1. Ковальчук Ю. О. Теорія валідизації вимірювань М.Т.Кейна як методологічна основа викладання розділу "Валідність" навчального курсу "Освітні вимірювання та моніторинг якості освіти" // Наукові записки. Серія "Психолого-педагогічні науки" / Ю. О. Ковальчук ; за заг. ред. проф. Є. І. Коваленко. – Ніжин : Видавництво НДУ ім. М.Гоголя, 2011. – № 7. – С. 8–12.
2. Ковальчук Ю. О. Валідність освітніх вимірювань / Ю. О. Ковальчук // Підготовка молодих викладачів та аспірантів в галузі освітніх вимірювань : тези міжнародного семінару молодих науковців (Кіровоград, Україна, 11–14 травня 2011 р.) – Кіровоград : РВВ КДПУ ім. В.Винниченка, 2011. – С. 34–35.
3. Toulmin S. E. The Uses of Argument (1958) / S. E. Toulmin. – 2nd ed. – 2003.
4. ECTS Users' Guide. – Luxembourg : Office for Official Publications of the European Communities, 2009.