МІЖНАРОДНА ТЕСТОВА КОМІСІЯ

Керівництво Міжнародної тестової комісії з великомасштабної оцінки лінгвістично та культурно відмінного населення

Версія 4.2

Варто цитувати як:

International Test Commission. (2018). ITC Guidelines for the Large-Scale Assessment of Linguistically and Culturally Diverse Populations. [www.InTestCom.org]

Подяка

Міжнародна тестова комісія (МТК) висловлює подяку Рене Лоулс (США) та Марії Єлєні Олівері (США) за розробку цього керівництва та участь у ролі головуючих комітету з даного проекту.

Міжнародна тестова комісія також дякує членам комітету за їхній внесок у проект. Серед них Аві Аллалуф (Ізраїль), Сіделл Карлтон (США), Томас Екес (Німеччина), Паула Елосуа (Іспанія), Молі Фолкнер-Бонд (США), Рональд Хамблтон (США), Драгош Іліеску (Румунія), Стівен Сіречі (США), Фонс ван де Вейвер (Нідерланди), Аліна фон Дав'є (США), та Ейпріл Зеніскі (США).

Міжнародна тестова комісія висловлює подяку членам Ради МТК, а також Кейті Вендлер та Роберту Міслеві, які надавали цінний зворотний зв'язок щодо попередніх версій цього керівництва.

Короткий огляд та сфера застосування

Дане керівництво описує міркування, що мають відношення до оцінювання учасників тестування в країнах або регіонах, що є лінгвістично та культурно відмінними. Методичні рекомендації було розроблено комітетом спеціалістів з метою інформування розробників тестів, психометристів, користувачів та адміністраторів тестувань щодо питань справедливості для забезпечення рівного та достовірного оцінювання лінгвістично та культурно відмінного населення. Вони мають бути застосовані до більшості чи майже всіх аспектів розробки, проведення, оцінювання та використання отриманих у результаті тестування оцінок. Ці рекомендації покликані доповнити інші вже існуючі професійні стандарти та керівництва з тестування та оцінювання. Таким чином, дані рекомендації зосереджено на типах адаптивного тесту та міркуваннях, якими необхідно керуватися в ході розробки, перегляду та інтерпретації даних та тестових балів, отриманих у результаті проведення тестів серед лінгвістично та культурно відмінного населення. Інші керівництва, такі як «Стандарти освітнього та психологічного тестування» (AERA, APA, & NCME, 2014), можуть також бути актуальними для тестування лінгвістично та культурно різноманітних груп населення.

Ці керівні принципи розроблені з метою інформування розробників тестів, психометристів та користувачів тестів про заходи, що мають бути вжиті для забезпечення об'єктивності оцінювання та порівнянності оцінок для підтримки значущих висновків у культурно та лінгвістично різноманітних контекстах. Вони розширюють вже існуючі керівництва МТК та інші професійні рекомендації (або стандарти), посилання на які вказані в кінці цього документу. Хоча керівництво в основному стосується великомасштабних оцінок, що проводяться в сфері освіти, його загальні принципи також можуть бути застосовані в інших умовах, таких як ліцензування, сертифікація та оцінювання володіння навичками (як такі, що перевіряються для отримання водійських прав). У невеликих масштабах або індивідуальних (клінічних) оцінюваннях можливі труднощі у реалізації даних рекомендацій. Окленд (2016) надає конкретні рекомендації щодо практичного застосування цих рекомендацій у ході проведення оцінювання осіб, які є іммігрантами та/або вивчають другу мову.

Існує велика чутливість щодо термінології, що використовується по відношенню до різних мов, якими спілкуються в країні або регіоні. Тому в цьому керівництві наскільки можливо обмежується використання таких термінів як мови меншин та більшості, або рідна та іноземна мова при позначенні різних мов, якими проводиться оцінювання в країні або регіоні. Деякі з численних факторів, що впливають на мовне розмаїття, наведені у вступі. У цьому керівництві посилання на мовні групи можуть також відноситися до груп з культурно та/або історично другою мовою (L2), базуючись на контексті стандартів. Така термінологія була обрана для забезпечення ефективності.

Глосарій рідко вживаної або технічної термінології можна знайти в кінці цього документу. Підкреслені терміни в цьому документі пов'язані з глосарієм. Щоб отримати доступ до цих визначень, ***натисніть Ctrl та гіперпосилання****.*

Зміст

Вступ

Важливим фактором вирішення питань справедливості та об'єктивності оцінювання є розгляд того, як задовольнити мовні потреби учасників тестування у лінгвістично відмінних країнах або регіонах. Подібні лінгвістичні контексти можуть бути результатом міграції (з економічних, соціальних, політичних або релігійних причин); інтересу до збереження та відновлення інших мов; минулої колонізації; та/або інших умов, що дозволяють людям переміщатися з одного регіону/країни до іншого регіону чи країни. По відношенню до населення з різноманітними лінгвістичними групами необхідні різні міри, особливо у випадках, коли рідна мова (мови) учасників тестування відрізняється від мови, якою навчають в школі, спілкуються в суспільстві або яка використовується в тесті. Одним зі складних моментів, які можуть виникнути, є ідентифікація домінуючої мови досліджуваного. У країнах, де існує більше однієї офіційної мови, можуть бути необхідні додаткові заходи.

Це керівництво розроблене з метою інформування розробників тестів, психометристів та користувачів тестів про питання рівності для підтримки справедливого та достовірного оцінювання лінгвістично чи культурно відмінних груп. Воно має бути застосовано до більшості чи майже всіх аспектів розробки, проведення, оцінювання та використання отриманих у результаті тестування оцінок, а також покликане доповнити інші вже існуючі професійні стандарти та керівництва. З огляду на те, що оцінювання використовуються для інформування щодо різноманітних рішень (деякі з яких є доленосними), керівництво охоплює міркування щодо широти життєвого циклу оцінки; тобто від її концептуалізації до реалізації та інтерпретації балів.

Фактори, що впливають на справедливу оцінку лінгвістично чи культурно відмінних груп населення

Центральним моментом у розробці або адаптації об'єктивних та достовірних оцінювань для лінгвістично чи культурно різноманітних груп населення є розгляд контекстних факторів, що впливають на процес відповіді учасників тестування. Індивідуальні особливості володіння мовою в учасників тестування можуть бути зумовлені відмінностями в тому, як саме у них проходив процес оволодіння та вивчення тієї чи іншої мови. Заслуговують на увагу й інші відмінності, що можуть бути пов'язані з соціальними аспектами для груп осіб, які беруть участь у тестуванні. Мова, якою користується установа, може відрізнятися, оскільки це стосується офіційного визнання або невизнання мови учасника тестування. Ці різноманітні фактори впливають на доступність ресурсів та навчальних матеріалів, ступінь підготовки вчителів з викладання певної мови та рівень уваги, що приділяється відновленню мови. Також існують відмінності у кодифікації мов. Наприклад, є мови, що існують в усній формі; тобто вони або не були кодифіковані взагалі, або були кодифіковані нещодавно, або ще перебувають в процесі кодифікації.

Правовий статус різних мов у межах країн

Деякі країни надають офіційного статусу декільком мовам. Це означає, що ці мови можуть використовуватись в державних установах. Навіть якщо мова не є широко вживаною, вона може мати правовий статус. Наприклад у Новій Зеландії є три офіційні мови: англійська,

маорі та новозеландська мова жестів. У Бельгії також три офіційні мови: французька, фламандська (нідерландська) і німецька. Іспанська мова є національною мовою Іспанії, проте країна має інші п'ять офіційних автономних регіональних мов: галісійська, баська, аранська, валенсійська та каталонська. Офіційний статус мови впливає на можливість забезпечення ресурсів, фінансової підтримки, створення/адаптації нових навчальних матеріалів, поширення мови на нові області та підготовку вчителів.

Мова інструкції

Важливим регулятором правильної роботи тесту є мова шкільного навчання. У ситуації, коли мова інструкції тесту не співпадає з рідною мовою (мовами) досліджуваного, важливо вирішити, яку мову тестування доречніше використовувати для того, щоб точно оцінити результати тестування та забезпечити валідність висновків отриманих на основі набраних респондентом балів. Це дає змогу розмежувати знання учасника тестування з конструкту, що оцінюється, та його рівень володіння мовою.

Вибір найбільш доречної мови тестування в таких випадках – непроста задача. З одного боку, виникає питання, чи доступний тест мовою, що є рідною для учасника тестування. З іншого боку, необхідно зрозуміти, чи володіє учасник тестування на достатньому рівні мовою, якою проводиться тест, щоб отриманий на основі балів висновок щодо його знання з оцінюваного змісту або конструкту був валідним та не викривлявся знанням мови тестування.

Виникають й інші ускладнення в ході прийняття рішення про те, яка мова має бути обрана для таких учасників тестування. Це має бути рідна мова учасника тестування, мова регіону, в якому проводиться тестування, чи, у випадку освітніх оцінювань, мова, якою відбувалося викладання предмету? Також виникають питання щодо того, хто саме повинен приймати це рішення та якими критеріями керуватися при їх прийнятті. Подібні рішення вимагають оцінки того, що є практично здійсненним та чи є наявними ресурси для втілення обраного рішення.

Дуже складна ситуація у виборі мови тестування виникає в країнах, що мають лінгвістично або культурно різноманітні групи населення. Для прикладу наведемо Південну Африку, в якій існує 11 офіційних мов, з яких не всі розглядаються аналогічно. Деякі з цих мов можуть отримати сильнішу підтримку в освіті та суспільному житті порівняно з іншими. Оскільки англійська є найбільш поширеною та вживаною мовою країни, це може зробити її кращим варіантом вибору мови тестування. Проте відмінності у якості освіти в школах по країні призводять до такої значущої розбіжності у володінні цією мовою серед учасників тестування, що використання англійської як основної мови тестування неможливе без попередньої перевірки рівню володіння нею у учасників тестування.

Інституційна підтримка

Поміж інших факторів, доступність ресурсів у країні залежить також від правового статусу мови, достатку країни та престижу різних мов, що співіснують в її межах. У деяких випадках можливе використання тестів з інших країн. Однак існують обмеження, що пов'язані з можливістю та доречністю використання оригінальних нормативних показників;

збереженням коректного визначення конструкта та відповідності навчальних програм у групах; або порівнянності балів для численних груп населення, що проходять тестування. Розгляд цих питань важливий, оскільки вони можуть вплинути на точність інтерпретацій отриманих балів. Подібні міркування можуть відноситись також до використання інструментів, що були розроблені однією мовою (наприклад, англійською) для однієї країни, а потім використовуються в інших країнах, в яких говорять "тією ж" мовою.

Кодифікація

Велика кількість мов кодифікована алфавітом або письмовим кодом, деякі ж з мов не кодифіковані. Наприклад, в Марокко є мови (арабська, французька), що мають офіційний статус і кодифіковані, але є й такі (наприклад, берберські мови), які існують лише у вербальній формі. Відмінності в лінгвістичній кодифікації або статусі підвищують складність створення відповідних адаптацій тестів для різних популяцій.

Соціальний статус і престиж мов

Лінгвістичні відмінності можуть існувати також у зв'язку з соціальним статусом та престижністю мов відносно основної мови країни чи регіону (у певних випадках на це впливає також політичне або культурне домінуюче становище), що разом можуть відображати різне ставлення до тієї чи іншої мови.

Мови в багатомовних суспільствах можуть відрізнятися за своїм соціальним статусом та престижем. Престиж мови (мов) означає рівень поваги до мови (мов) або діалекту в порівнянні з іншими мовами або діалектами певної мовної спільноти. Урахування соціолінгвістичного контексту мови може допомогти освітнім органам обрати мову оцінювання, а також інтерпретувати будь-які можливі відмінності у балах серед представників різних лінгвістичних груп.

Звичайно, не існує загального та універсального контексту, що чітко визначав би характеристики лінгвістично чи культурно відмінних груп населення, оскільки вони відрізняються в різних країнах і в межах країн. Тому прості чи шаблонні рекомендації в подібних ситуаціях мають обмежену цінність.

Різне використання мов

Додаткові труднощі з деякими мовами (наприклад, з арабською та китайською мовами) виникають через особливості їх використання, тобто через різницю в усному (розмовна арабська) та кодифікованому використанні мови (писемна арабська). У випадку з арабською мовою існують регіональні відмінності у використанні слів. У випадку китайської мови, населення розмовляє декількома мовами, проте всі вони використовують одну й ту саму письмову систему базовану на символах. Таким чином слова у розмовній мові не завжди ідентичні письмовій формі.

Інші відмінності можуть виникати на регіональному рівні. Наприклад, у Канаді можуть бути учасники тестування, які розмовляють французькою або англійською мовою залежно від того, в якій спільноті вони проживають – у англомовній чи франкомовній. Деякі учасники тестування можуть бути корінними жителями Канади, а інші – іммігрантами, що нещодавно переїхали. Можуть бути й такі учасники тестування, які протягом декількох років

знаходяться в стані переїзду та їх процес вивчення мови, якою написана інструкція до тестування, є, відповідно, уривчастим. У зв'язку з високим рівнем імміграції в Канаді регіональні відмінності також можуть виникати серед носіїв французької мови, наприклад через те, що вони походять з різних франкомовних країн, деякі з яких можуть мати різні варіації французької мови, в тому числі креольську.

Наявність ресурсів для адаптації тестів

Наявність ресурсів залежить, серед інших факторів, від достатку країни, офіційного статусу мов іммігрантів (в межах регіону або країни), а також від того, наскільки добре організоване іммігрантське населення. У деяких випадках можуть бути використані тести з інших країн. Однак використання таких тестів (та їх нормативних показників) може бути проблематичним, оскільки інструменти, як правило, не призначені для використання за межами країни, в якій вони були розроблені, і можуть вимірювати різний зміст або конструкти, або мати невідповідні дані, з якими мають бути порівняні результати тестування. Подібні складнощі можуть виникнути у використанні розроблених англійською мовою інструментів у англомовних країнах, в яких може відрізнятися використання конкретної термінології або фраз. Виправлення або коригування цих відмінностей на основі наявних ресурсів може бути нездійсненним.

Керівні принципи

Принцип 1: Розробка та адаптація тесту

Адаптація існуючих тестів для лінгвістично чи культурно відмінного населення[1]

1.1. Розробники/видавці тестів повинні враховувати лінгвістичні відмінності між вихідними та цільовими мовами та культурами (граматичні, синтаксичні, семантичні, лексичні та ін.) при адаптації тестів або інших інструментів для досліджуваних з цільовими мовами, щоб зробити форми тесту настільки психометрично порівнянними, наскільки це можливо. Особливу увагу необхідно приділяти ситуаціям, коли мови належать до різних мовних сімей.

1.1.1. До розробки питань та тесту, який буде адаптований, повинні бути залучені особи із різних мовних груп, оскільки вони найкраще можуть виявити будь-які лінгвістичні неточності чи складнощі, що можуть виникнути в ході перекладу, та внести пропозиції щодо того, як їх уникнути.

1.1.2. Слід враховувати культурні аспекти при перекладі компонентів тесту (питань, шкал, вказівок тощо) і докладати зусиль для того, щоб адаптувати тест з його оригінального виду до цільового не тільки лінгвістично, але й культурально.

1.2. У разі необхідності, адаптуйте формулювання питань тесту (з мови оригіналу) для учасників тестування, якщо це не змінює конструкту.

1.2.1. Тест, перекладений цільовою мовою повинен бути аналогічним за довжиною до оригінального тесту, і кожне питання має містити таку ж кількість варіантів вибору, як і питання в оригінальному тесті, що адаптується.

1.2.2. Питання, сформульовані цільовою мовою повинні мати той самий реєстр, що й питання оригінального тесту, такий самий рівень складності, і не включати конотації, які відсутні в оригіналі.

1.2.3. Оскільки дослівний переклад питань та елементів тесту можуть не мати сенсу цільовою мовою, переклади мають передати синонімічні ідеї, пов'язані з конструктом, не змінюючи складності питання.

1.3. Якщо тест розроблений для декількох культур та/або мов, або адаптований для певної цільової культури та/або мови, враховуйте, що формат питань, стимулів, критеріїв оцінки та інструкцій з тестування мають бути в однаковій мірі знайомі для всіх цільових груп.

1.3.1. Процедури створення вибірки повинні бути якомога більш схожими для форм тесту різних мов та/або культур з метою запобігання упередженості в ході аналізу еквівалентності.

[1] Для ознайомлення з оглядом принципів перекладу та адаптації на високому рівні див. Керівництво МТК з перекладу та адаптації (2-е видання), www.intestcom.org

1.3.2. Різний рівень ознайомленості зі стимульними матеріалами, відмінні стилі відповіді або різний рівень соціальної бажаності для лінгвістично і/або культурно відмінного населення можуть призвести до упередженості інструменту. Ці відмінності мають бути досліджені під час аналізу порівнянності.

1.4. Коли тест адаптовано, переконайтеся, що розміщення на сторінці таких елементів, як фотографії та номери сторінок, не перешкоджає читабельності тексту. Перегляньте всі цифри (зображення) у адаптованих питаннях на придатність для всіх мовних груп.

1.4.1. Зовнішній вигляд та розміщення елементів тесту, перекладеного на цільову мову, повинні бути максимально схожими на такі у оригінальному тесті. Наприклад, учасники тестування, що проводиться цільовою мовою, не мають бути в невигідному положенні через те, що їм необхідно перегорнути сторінку або прокрутити документ, щоб побачити текст, в той час як для учасників тесту на вихідній мові весь текст з'являється на одній сторінці. Крім того, для мов, що читаються зліва-направо переконайтеся, що зображення відзеркалено (або не відзеркалено) залежно від норм читання в країні.

1.5. Всі адаптовані тести повинні оцінюватися на предмет точності рецензентами (які володіють не тільки мовою оригіналу та цільовою мовою тесту, але й розбираються в обох культурах), щоб забезпечити точність конструкту та належний переклад. Будь-які зроблені адаптації повинні бути задокументовані та надані користувачу тесту.

Зрозумілість формату питань та розробка тестових інструкцій

1.6. Розробляйте тестові інструкції з максимальною чіткістю та зрозумілістю (використовуйте просту та зрозумілу мову).

1.6.1. Представляйте інструкції до тестування, використовуючи різні форми подання матеріалу (наприклад, усну та письмову форму); де це можливо, надавайте інструкцію домінуючою мовою учасників тестування, окрім випадків коли тест оцінює володіння мовою.

1.7. Розробники/видавці тестів повинні надавати докази (такі як редакційні огляди або огляди з перевірки справедливості) того, що мова, яка використовується в інструкціях, вказівках та питаннях тесту, є зрозумілою для тих, хто цей тест використовує та проходить.

1.8. Не припускайте, що учасники тестування L2 мають попередній досвід роботи з даними типами завдань або питань. Оцінюйте ознайомленість з форматом питань, щоб переконатися в тому, що вони є придатними для всіх учасників тесту, незалежно від їх мовної групи. Надавати перевагу слід таким форматам питань, які підходять для всіх груп населення, а не тим, які необхідно змінювати для різних груп.

Розробка питань та огляд

1.9. При виборі тем для питань, уникайте таких, які можуть вважатися образливими, принизливими, чи такими, що відокремлюють, або можуть викликати емоційну реакцію членів будь-якої з лінгвістично чи культурно різноманітних груп населення, оскільки це може створити упередження, що не відноситься до конструкту.

1.10. Розробляйте питання тесту та текстові фрагменти таким чином, щоби вони містили доступну для всіх мовних та культурних груп лексику. Необхідно використовувати мову, що не містить регіональну або специфічну лексику. Крім того, уникайте використання слів з кількома значеннями або інших занадто складних слів, які не є частиною конструкту, що оцінюється.

1.10.1. В ситуації, коли питання пишуться для кількох мовних груп, необхідно проконсультуватися з носіями мови із кожної мовної групи, щоб уникнути проблемної термінології, як наприклад, регіональної або специфічної для певної групи населення.

1.11. У випадках, коли це можливо, уникайте використання неоднозначної мови з вихідної мовної версії тесту, наприклад, використання скорочених слів у підказках, оскільки може бути складно розробити адаптації таких термінів для різних мов.

1.12. Коли це не входить до оцінюваного конструкту, розробляйте питання з простою структурою речення. Краще надати перевагу декільком більш коротким реченням, ніж одній складній фразі.

1.13. Коли це можливо, розробляйте питання таким чином, щоб в них що відображалися контексти та сценарії загальні для всіх мовних груп і різного населення.

1.14. Коли це не є частиною оцінюваного конструкту, уникайте посилання на історичний контекст і назви, що можуть бути добре відомі одним культурам та не відомі іншим.

1.15. При розробці питань тесту притримуйтесь мінімального рівню вимог до знання мови, необхідного для оцінки конструкту.

1.16. Уникайте використання нерелевантних конструкту назв виробів, осіб, географічних об'єктів, урядових назв, свят, вимірювальних систем та грошових одиниць, які можуть бути релевантними або більш знайомими лише для деяких культурних/мовних груп.

1.17. При адаптації питань на версії цільових мов, зверніть особливу увагу на виявлення та уникнення формулювань, що можуть мати різні значення для різних мовних груп.

1.18. Повинні бути надані огляди експертів з кожної мовної групи для того, щоб гарантувати те, що питання охоплюють необхідний конструкт для всіх мовних груп. Експерти, які переглядають або обирають питання, повинні бути добре обізнані з культурою різних мовних груп і вільно володіти мовою питань, які вони розглядають. В ідеалі, ці експерти повинні належати до цільової культури та вільно володіти цільовою мовою.

1.19. Питання, рейтингові шкали та тестові матеріали мають бути переглянуті на наявність елементів (наприклад, історичних подій, ситуацій, малюнків, кольорів), до яких члени різних мовних груп можуть бути чутливими або про які вони необізнані. Рекомендовано залучення експертів з лінгвістичної/культурної оцінки для проведення цих оглядів на початковій стадії розробки питань.

1.20. Коли це можливо, фонові демографічні питання, що мають бути розроблені для тесту, мають чітко та з достатньою деталізацією запитувати про мовний фон учасників тестування для того, щоб дозволити надалі провести змістовний аналіз на рівні групи.

Випробування питань

1.21. Якщо це можливо, проводьте випробування питань або когнітивні інтерв'ю з учасниками тестування з усіх лінгвістичних груп, щоб впевнитись у доречності питань для кожної мовної групи, і визначити, чи взаємодіють учасники тестування з кожної мовної групи з питаннями бажаним чином.

1.22. Якщо питання в тестуванні не потребує використання мови, як наприклад рівняння або стимульне зображення, і його адаптація не потрібна, надайте емпіричні докази щодо порівнянності між групами учасників, що вивчають першу мову (L1) та другу мову (L2).

1.23. Будь-які питань, при відповіді на які учасники тестування з деяких лінгвістичних груп не використовують передбачувані процеси вирішення таких завдань, мають бути оцінені (разом з візуальними підказками, інструкціями, вказівками тощо) для визначення можливих змін, що можуть зробити ці питання більш зрозумілими та зручними для учасників тестування.

1.24. Якщо є достатня кількість даних по всіх мовних групах, проводьте статистичний аналіз, щоб переконатися, що питання працюють однаково для різних лінгвістичних груп.

Принцип 2: Валідність, надійність та справедливість

Валідність

2.1. Коли адаптована версія тесту використовується як пристосування для оцінки будь-яких учасників тестування, необхідно провести оцінку порівнянності адаптованого варіанту з оригінальною версією тесту.

2.1.1. Дослідження валідності мають бути проведені, щоб впевнитися в тому, що адаптована версія тесту вимірює конструкт(-и), передбачуваний(-ні) призначенням тесту.

2.1.2. Слід розглядати паралельну/поєднуючу модель тесту між версіями, включаючи якірні питання на рівні проектування тесту та долучаючи звичайних осіб під час попереднього випробування або проведення.

2.2. Переконайтеся, що відношення між тестовими балами та іншими змінними є порівнянними між усіма мовними та культурними групами.

2.2.1. Докази валідності, засновані на співвідношенні тестових балів з іншими змінними, можуть надавати важливі дані щодо того, наскільки добре результати тестів відповідають передбачуваному призначенню. В ході тестування лінгвістично чи культурно відмінних груп населення слід дослідити ступінь, до якого ці відношення дотримуються для підгруп учасників тестування, що вирізняються мовною або культурною різноманітністю. Будь-які розбіжності, виявлені в кореляції або прогнозі, можуть вимагати подальшого дослідження та/або документації для розслідування можливих непередбачених наслідків.

2.3. Якщо для учасників тестування з різних лінгвістичних груп розробляються різні версії питань, ці зміни мають бути задокументовані, а інваріантність їх психометричних характеристик має бути включена в документацію, включаючи будь-який вплив змін на інтерпретацію балів.

2.4. Якщо інтерпретація балів може бути різною для різних мовних груп (наприклад, окремі таблиці норм для груп, визначених за країною або мовою), слід надати обґрунтування для дозволу таких варіацій і задокументувати їх вплив на інтерпретацію та використання результатів тесту.

2.4.1. В разі якщо забезпечення повної інваріантності між різними мовними формами тесту неможливе, часткова інваріантність є прийнятним компромісом. Часткова інваріантність встановлює незмінність для підмножин питань, а не всіх питань тесту. Якщо на таких субтестах проводиться аналіз, інваріантність може бути підтверджена, але при цьому охоплення тестом змісту може змінюватись, і тому значущість балів може також відрізнятися. У цьому випадку необхідно надати документацію про такі зміни.

2.5. Оцініть інваріантність внутрішньої факторної структури оцінювання для мовних груп L1 та L2.

2.6. Якщо виявлено, що учасники тестування, які проходять тест різними мовами, складають його з різними показниками відносно один одного, підкріплюйте цей факт іншими формами емпіричних доказів, щоб показати, що такі відмінні показники не пов'язані з упередженістю конструкції тесту або оцінювання.

2.6.1. Якщо результати тестування серед різних лінгвістичних груп виявляються непорівнянними, надайте докази того, що подібні відмінності не матимуть негативного впливу на результати тесту. Якщо ж це має негативний вплив, приведіть вагомі докази того, що тест служить призначеній меті без незапланованих негативних наслідків.

2.7. Сам тест та спосіб проведення тестування мають належним чином враховувати діапазон здібностей всіх учасників тесту, включаючи різні лінгвістичні групи учасників тестування.

2.7.1. Для належного оцінювання діапазону здібностей усіх осіб, які беруть участь у тестуванні, слід розглянути варіативність розподілу здібностей серед різних лінгвістичних груп досліджуваних.

2.7.2. При виборі способу проведення тестування (адаптоване під комп'ютери тестування, багатоступеневе адаптивне тестування, модульне або лінійне тестування) слід враховувати діапазон можливостей учасників тестування мовної групи L2 та їх ознайомленість з такими форматами проведення тестування.

2.7.3. Якщо відмінність балів між різними лінгвістичними групами велика, розгляньте варіант використання адаптивного тестування (замість лінійного) для ефективного підвищення точності оцінювання усіх осіб, які беруть участь у тестуванні.

Релевантність конструкту

2.8. Релевантність (відповідність) конструкту, виміряна як для учасників тестування групи L1, так і для L2, має бути задокументована. Така документація повинна включати обґрунтовані аргументи, зроблені з соціокультурної точки зору, та аргументи, що базуються на емпіричних даних (тобто докази того, що валідність інтерпретації балів є рівноцінною для усіх мовних та культурних груп).

2.9. Коли докази релевантності конструкту для лінгвістично або культурно відмінних учасників тестування базуються на експертних оцінках, характеристики вибірки учасників або експертних суддів мають бути задокументовані.

2.10. Для виявлення рівню володіння мовою учасників тестування (L2) використовуйте окреме тестування саме для оцінювання мови. Якщо це можливо, щорічно проводьте цей тест (з використанням різних паралельних тестових форм), оскільки рівень володіння мовою учасників тестування може змінюватися з року в рік.

2.10.1. Визначте рівень знання найбільш підходящої мови тесту у осіб, що беруть участь у тестуванні, якщо тільки саме знання мови не виступає досліджуваним конструктом.

Надійність

2.11. Переконайтеся, що бали тестування відповідають прийнятним критеріям надійності для кожної мовної групи, як в абсолютному сенсі, так і відносно груп населення, що були задіяні у вихідній формі тесту.

2.11.1. У разі необхідності проводьте аналіз надійності, щоб забезпечити прийнятний мінімальний стандарт надійності для всіх мовних груп. Приклади таких аналізів включають коефіцієнт альфа, надійність потворного тестування (ретест), функції тестової інформації, узгодженість класифікації/прийняття рішень, аналіз стандартних помилок вимірювання та умовні стандартні похибки у порівнянні з показниками межового балу.

Справедливість

2.12. Проводьте перевірку справедливості для всіх питань та елементів тестування (включаючи підказки, інструкції, зображення, вказівки) зосереджуючись на лінгвістично чи культурно різноманітних групах учасників тестування.

2.12.1. Залучайте представників усіх мовних та культурних груп в експертні комісії з мови, що використовуються для перевірки справедливості.

2.12.2. Наскільки це можливо, всі тестові матеріали мають бути перевірені на відсутність в них:

- Образливих або надто узагальнених зображення різних мовних груп;
- Зображень або посилань, які з великою ймовірністю будуть незнайомі для учасників тестування з різних мовних груп і не мають безпосереднього відношення до оцінюваного конструкту;
- Зображень або фраз, які є образливими для інших культур або релігій;
- Мови, зображень або вмісту, що, ймовірно, будуть невиправдано вигідними або невигідними для учасників тестування з різних мовних груп.

2.13. Залучайте учасників тестування з усіх лінгвістичних і культурних груп до нормативної групи (якщо оцінки будуть нормованими) або експертів з усіх мовних та культурних груп до комісій (якщо оцінки будуть критеріальними) для забезпечення представлення всіх мовних і культурних груп при визначенні стандартів виконання тесту.

2.13.1. Матеріали, що надаються особам, що проводять тестування, мають містити опис стандартних умов проведення та нормативні дані, якщо такі є. Такі дані мають містити інформацію про демографічні показники нормативних груп та дату, коли оцінювання було нормоване.

Порівнянність балів

2.14. Для адаптованих тестів проводьте дослідження порівнянності балів з метою вивчення ступеня інваріантності тестових балів для обох варіантів тесту.

2.14.1. Якщо докази достовірності вказують на те, що бали адаптованого і оригінального тестів є порівнянними, їх слід розглядати так само, як і всі інші бали.

2.14.2. Якщо докази достовірності вказують на те, що бали адаптованого і оригінального тестів не є порівнянними, то (а) необхідно переглянути кроки, що були зроблені для забезпечення порівнянності, і (б) необхідно переглянути процедури адаптації тесту. Крім того, експерти повинні розглянути неспівставні питання, щоб визначити, чи вони можуть бути додатково адаптовані для встановлення порівнянності з оригінальним тестом.

2.14.3. Якщо докази достовірності вказують на те, що бали адаптованого і оригінального тестів не є порівнянними, користувачі тесту мають бути поінформовані про непорівнянність за допомогою легкодоступної документації.

2.15. Надайте чітке обґрунтування та підтверджуючі докази, щоб довести, що бали між формами тестування, перекладеного різними мовами, є порівнянними. Це включає в себе вказівки, чітко визначену підготовку оцінювачів та використання статистичних моделей (таких як теорія відповіді на питання) для оцінки порівнянності.

2.15.1. Щоб підтвердити порівнянність балів між різними лінгвістичними та культурними групами, надайте докази вимірювання інваріантності балів, включаючи диференційний аналіз функціонування питань (DIF), якщо є вибірка достатнього розміру.[2]

2.15.2. У відповідних випадках надайте детальну технічну інформацію про метод, обраний для порівняння/зв'язування балів між двома версіями тесту на різних мовах.

2.16. Якщо бали тесту мають бути нормованими, різні лінгвістичні групи повинні бути представлені в нормативній групі в мовній версії цього тесту. Нормування має ґрунтуватися на населенні та лінгвістичному різноманітті регіону, де проводитиметься тест.

2.17. Якщо виявляються суттєві та систематичні відмінності у порівнянні балів між мовними групами, слід проводити дослідження (наприклад, лінгвістичний/культурний аналіз), щоб забезпечити те, що ці відмінності не призведуть до розбіжностей у оцінках, які б ставили у невигідне положення будь-яку лінгвістичну групу учасників тестування.

Дослідження джерел відмінного функціонування питань

2.18. Проводьте диференційний аналіз функціонування питань (DIF) для кожного питання відповідно до типу питання та оцінюваного конструкту, щоб гарантувати, що учасники тестування з різних лінгвістичних груп взаємодіють з питаннями так само, як і учасники тестування з контрольної групи.

2.18.1. Якщо розміри вибірки є достатньо великими, визначте підгрупи, що вирізняються в кожній лінгвістичній групі та розгляньте можливість проведення DIF для кожної підгрупи.

[2] Для отримання додаткової інформації про різні процедури, які можна використовувати для аналізу DIF, див. розділ «Керівництво з підтвердження» у Керівництві МТК з перекладу та адаптації тестів (друге видання), www.intestcom.org.

2.19. Оцініть зміст питань для джерел DIF, щоб дослідити можливі джерела нерелевантних конструкту відмінностей.

2.19.1. Коли це можливо, назначте експертам з лінгвістичної / культурної оцінки проведення перевірки питань, що позначені для DIF.

2.19.2. Якщо виявлено, що джерело DIF не має відношення до оцінюваного конструкту, розгляньте можливість перегляду або вилучення питання.

Принцип 3: Оцінка есе та інших відкритих відповідей на питання

Розробка, проектування та адаптація вказівок щодо інтерпретації балів

3.1. Розробка вказівок щодо інтерпретації результатів тестування має відображати мовну різноманітність цільового населення, а також мету тестування.

3.1.1. Вказівки щодо інтерпретації балів повинні бути розроблені таким чином, щоб вони не ставили у невигідне положення осіб, для яких мова тесту не є рідною, чи учасників тестування, які інтерпретують питання не так як інші на основі їх специфічного культурного контексту. Наприклад, якщо учасники тестування відповідають на короткі відкриті питання пов'язані з наукою, їх рівень володіння мовою не має впливати на бали, що вони отримають в результаті тестування, тільки якщо це не заважає зрозумілості їх відповіді.

3.1.2. Мають бути створені примітки для тих, хто оцінює результати тестування, щоби не допустити існування штрафних санкцій для особливих способів використання мови, а обмеженість володіння мовою або культурні відмінності помилково не сприймалися як обмежене знання з конструкту.

3.2. Різні схеми оцінювання можуть впливати на учасників тестування з різних мовних груп неоднаково.

3.2.1. Коли можуть бути наявними відмінності в конкретних аспектах або сторонах оцінюваного конструкту між лінгвістичними групами, користуйтеся способом оцінювання, що розрізняє ці аспекти (аналітична оцінка/ оцінка ознак).

3.2.2. При використанні загального показника для оцінювання взаємопов'язаних навичок переконайтеся, що загальний бал результатів тестування порівнянний між мовними групами.

3.3. Беручи до уваги потенційну неоднорідність відповідей в різних лінгвістичних групах, проводьте попередню перевірку розроблених критеріїв інтерпретації результатів на репрезентативній вибірці з загальної популяції, включаючи в тому числі респондентів з усіх мовних груп.

3.3.1. Проводьте аналіз, щоб отримати уявлення про ефективність та корисність кожної з категорій рейтингової шкали (наприклад, рівень соціальної бажаності). За можливістю, проводьте цей аналіз з усіма мовними групами, щоб з'ясувати, чи виникають розбіжності між групами при використанні категорій рейтингової шкали.

3.3.2. Якщо різниця у використанні вказівок для деяких лінгвістичних груп має вплив на їхні результати, може бути доречним або необхідним перерахування балів по питаннях після перегляду вказівок та/або повторного орієнтування та інструктування оцінювачів. Метою перерахунку балів є не штучне збільшення або зменшення балів конкретної мовної групи, а з'ясування того, чи мають відношення до конструкту існуючі відмінності.

Матеріали для виставлення балів

3.4. Надайте примітки та еталонні приклади відповідей для всіх мовних груп, щоб описати потенційно різні стилістичні схеми написання, виділивши з них ті, які могли б призвести до оцінки відповідей нижчим балом по шкалі через причини, що не мають відношення до конструкту.

3.5. Всі оцінки мають проводитися анонімно. Довідкова інформація про осіб, які беруть участь у тестуванні, не повинна з'являтися на матеріалі, що буде оцінюватися, включаючи ім'я, країну походження, вік, стать, мовний фон, або етнічну чи культурну приналежність учасників тестування.

3.6. Коли бали виставляються кількома оцінювачами, переконайтеся, що їх оцінки не є видимими та не можуть бути ідентифіковані будь-яким чином іншими оцінювачами.

Відбір та навчання оцінювачів

3.7. Чітко визначте кваліфікації та характеристики, що мають бути притаманні новим оцінювачам, та обирайте їх на основі цих даних. В ідеалі, оцінювачі повинні мати попередній досвід оцінювання широкого кола виконання завдань учасниками тестування з різних мовних груп.

3.7.1. При використанні оцінювання конкретних завдань або оцінювання ознак (аналітичне оцінювання), коли це можливо, залучайте оцінювачів, які знаються на різних лінгвістичних групах.

3.8. Як група, оцінювачі повинні представляти широкий спектр демографічних, регіональних, змістовних та професійних характеристик, і, наскільки можливо, включати в себе членів лінгвістичних груп, які можуть вирішити проблеми розбіжностей в результатах тестування, що в свою чергу можуть надавати переваги або позбавляти їх учасників тестування з груп L2 через причини, що не мають відношення до вимірюваного конструкту. Ці оцінювачі можуть також допомогти правильно зорієнтувати інших оцінювачів, щоб забезпечити справедливість оцінювання для всіх груп населення.

3.9. Щоб забезпечити ефективний, чітко організований процес оцінювання, висококваліфіковані оцінювачі, які ознайомлені з відповідями від усіх мовних груп, повинні проводити та контролювати оцінювання, як керівники або голови комісій. Ці керівники несуть відповідальність за контроль роботи інших оцінювачів та за забезпечення відповідності оцінювання вказівкам.

3.10. Надавайте оцінювачам достатньо велику і різноманітну вибірку відповідей, що є нетиповими для цільового населення, включаючи еталонні приклади відповідей від усіх мовних груп.

Еталонні відповіді та межі діапазону (rangefinders)

3.11. Оцінювачі, що представляють різні мовні групи, повинні мати чітко визначені критерії, для оцінювання відповідей, аналізу контексту питань та вирішення розбіжностей між рішеннями оцінювачів. Ці критерії можуть бути використані при

підготовці та повторному навчанні інших оцінювачів. Використовуйте попередньо оцінені відповіді (еталони), щоб наочно представити кожну категорію рейтингової шкали або опис рівня у вказівках до оцінювання, включаючи відповіді учасників тестування з кожної мовної групи. Використовуйте ці еталонні відповіді щоб оцінити співвідношення балів оцінювачів з заданими критеріями, і посвідчити факт того, що робота оцінювачів успішно налаштована на еталонні відповіді.

3.11.1. Використовуйте межі діапазону, щоб допомогти оцінювачам узгоджено визначити інтервали категорій та позначити відповіді в областях рейтингової шкали, які є важливими для кращого розуміння розрізнення балів, особливо коли йдеться про учасників тестування з різних мовних груп.

3.12. Перевірте, чи є якісь розбіжності, що виникають у балах оцінювачів, присвоєних відповідям учасників тестування з кожної мовної групи, і обговоріть ці розбіжності, з'ясувавши, з яких причин вони виникають – пов'язаних чи непов'язаних з конструктом.

План оцінювання

3.13. Коли є група кваліфікованих та сертифікованих оцінювачів (кожен з досвідом оцінювання завдань учасників тестування щонайменше з однієї лінгвістичної групи), необхідно прийняти рішення щодо кількості оцінювачів, що будуть задіяні в процесах оцінювання. Незалежно від кількості оцінювачів, назначених на одну відповідь, оцінювачі повинні подавати незалежні оцінки, щоб уникнути небажаних наслідків, наприклад таких, як обговорення двома або більшою кількістю оцінювачів виставлених балів або імітація ними один одного.

3.14. У плані виставлення балів враховуйте наступні обмеження: графік, бюджет, важливість результатів тестування для осіб, які беруть в ньому участь (доленосні та рішення що мають менше значення), необхідний рівень надійності та проект виставлення балів (включаючи спосіб, у який оцінювачі призначаються учасникам тестування, завданням і відповідям). Наприклад, у ситуації коли оцінювання складається з декількох завдань, надійність буде вищою тоді, коли різні оцінювачі будуть виставляти оцінки результатам випробуваного, ніж тоді, коли один і той же оцінювач виставляє оцінки кожному завданню учасника тестування.

3.15. Розробляйте такий план оцінювання, що буде вигідним як по часовим, так і по фінансовим затратам, і одночасно дозволятиме керівнику оцінювання порівняти всіх оцінювачів, учасників тестування і завдання в межах однієї системи відліку.

3.16. Прагніть до створення такого плану оцінювання, який пов'язує оцінювачів, учасників тестування, критерії та завдання, враховуючи мовне різноманіття. Система зв'язків є обов'язковою для врахування, і позначається, наприклад, на відмінностях у рівні суворості або поблажливості, які кожен окремий оцінювач демонструє при присвоєнні оцінок учасникам тестування.

3.17. При оцінюванні відповідей, ставте в пари носіїв різних мов, щоб компенсувати будь-які потенційні упередження, пов'язані з впливом особливостей сприймання або точки зору кожного оцінювача.

Надійність та узгодженість між оцінювачами

3.18. Використовуйте показники надійності та узгодженості між оцінювачами для визначення міри, якою оцінювачі не погоджуються один з одним, щоб надати докази загального успіху процедур підготовки оцінювачів для всіх мовних груп.

3.19. Обчисліть принаймні два різні статистичні показники узгодженості між оцінювачами: індекс консенсусу, що вказує наскільки схожими (чи однаковими), є оцінки, що оцінювачі виставляють одним і тим самим відповідям (наприклад, відсоток повних або часткових збігів у оцінках) і другий індекс консенсусу, що вказує на ступінь, до якої оцінювачі послідовно ранжують відповіді випробувачів (наприклад, використовуючи коефіцієнт кореляції Пірсона).

3.20. У випадку доленосних рішень, вимоги до надійності є особливо жорсткими, тому залучайте принаймні двох незалежних оцінювачів для прийняття рішень щодо остаточної оцінки.

3.21. Порівняйте показники надійності та узгодженості між оцінювачами з робочими показниками (якщо два або більше оцінювачів надають оцінки для одного і того ж набору результатів) або обчисліть показники між робочими показниками та оцінками, наданими експертами або головами комісій з оцінювання.

Моніторинг та оцінка роботи оцінювачів

3.22. Регулярно стежте за балами, що виставляються оцінювачами, щоб підтримувати послідовність і точність процесу оцінювання, особливо в ситуаціях з доленосною значущістю результатів оцінювання для респондентів з різнорідних груп населення.

3.22.1. Керівники або голови комісій з оцінювання повинні застосовувати процедури перевірки якості read-behind або read-ahead, якщо такі є, зокрема в програмах онлайн-оцінки.

3.22.2. Використовуйте процедури read-ahead для визначення схожості та розбіжностей між балами оцінювачів та експертними оцінками.

3.22.3. Включайте результати процедур read-behind чи read-ahead в звіти про якість роботи окремих оцінювачів.

Повторне налаштування та перепідготовка оцінювачів

3.23. У разі значних відхилень від очікувань якості або стандартів, встановлених керівником оцінювання, проводьте перепідготовку або повторне налаштування оцінювачів, що виявляють неприпустимо низьку якість оцінки, використовуючи нові набори практичних відповідей, еталонних відповідей та межі діапазону.

3.24. Повторно призначайте оцінювачів, що пройшли перепідготовку, на оцінювання лише в тому випадку, якщо перевірки якості свідчать про достатньо високий рівень узгодженості з іншими оцінювачами та з керівниками оцінювання.

Статистика та зворотний зв'язок оцінювачів

3.25. У ході сесій оцінювання регулярно збирайте та аналізуйте інформацію, надану оцінювачами (тобто цілісні оцінки, аналітичні бали по субшкалах, загальні оцінки, використання категорій тощо) для отримання статистики оцінювачів.

3.26. Використовуйте статистичні дані, такі як засоби виставлення балів, стандартні відхилення оцінювачів, частоти оцінок або категорій шкали, а також узгодження з іншими оцінювачами, керівниками або головами комісій з оцінювання, щоб дати зворотній зв'язок окремим оцінювачам щодо їх процесу виставлення балів, наприклад, у випадках занадто поблажливого або жорсткого оцінювання.

3.27. Розрахуйте статистичні дані узгодженості та коефіцієнти надійності між оцінювачами, щоб надати докази того, наскільки кожен оцінювач відхиляється від інших, або навпаки, узгоджується з ними.

Підхід до вирішення розбіжностей в оцінюванні

3.28. Коли два або більше оцінювачів надають відмінні оцінки для одного і того ж набору відповідей учасника тестування, має бути точно визначений та використаний метод вирішення розбіжностей між оцінювачами для виставлення єдиної оцінки кожному учаснику тестування.

3.28.1. Методи вирішення включають усереднення двох балів (середнє значення), включення оцінки третього оцінювача під час винесення рішення (метод паритету) або заміну обох початкових балів оцінкою експертного судді (експертний метод). Вибір конкретного методу буде залежати від часу і бюджету, а також від наявності досвідчених експертів з оцінювання.

3.28.2. Оцінювачі, які представляють різні лінгвістичні групи, можуть вирішити можливі розбіжності в оцінках, що можуть ставити у вигідне або невигідне положення учасників тестування з мовних груп L2, через не пов’язані з конструктом причини.

Результати вимірювань та аналіз ефекту оцінювача

3.29. Якщо це можливо, спирайтеся на досвід психометричного моделювання балів, що спостерігаються, щоб ретельно відстежувати, аналізувати та оцінювати процес виставлення балів в ході оцінювання усіх мовних груп.

3.29.1. Вивчіть різні джерела помилок вимірювання (як-то відмінності діалектів, різні формати завдань або різні мовні версії одного і того ж тесту). Наприклад, використовуйте теорію узагальнення, щоб оцінити величину вкладу кожного з цих джерел і надати стратегію підвищення надійності тесту або оцінювання.

3.29.2. Оцініть міру спроможності кожного учасника тестування наскільки можливо незалежно від особливостей тесту або ситуації оцінювання, наприклад, щоб

компенсувати такі ефекти оцінювачів, як відмінності у поблажливості або жорсткості оцінювання будь-якої мовної групи. Для цієї мети використовуйте такі методи, як наприклад багатоаспектне вимірювання Раша (Many-Facet Rasch Measurement, MFRM).

3.30. Вивчіть потенційні відмінності, що можуть виникнути між різними лінгвістичними групами внаслідок відмінностей у процесах оцінювання (автоматизоване або з залученням людей).

3.30.1. Переконайтеся в тому, що учасники тестування з різних лінгвістичних груп не втрачають бали за відмінності у стилях написання, які можуть відрізнятися від референтної популяції з причин, що не є важливими для оцінюваного конструкту.

Принцип 4: Процедури проведення тестів та інструкції для пристосувань

Проведення тесту

4.1. Керівництво з проведення тесту має визначати всі аспекти проведення тесту, які потребують уважного вивчення для нових мовних чи культурних груп. Це керівництво має бути написане тією мовою, якою проводиться тест.

4.1.1. Наглядачі тестування повинні дослівно зачитувати будь-які тексти, що надаються для проведення тесту мовою тестування або мовою кожної з груп, що його проходять.

4.2. Коли це можливо, проводьте тестування мовою, якою учасники тестування володіють найкраще, якщо тільки саме знання мови не виступає конструктом, що оцінюється.

4.3. Якщо будь-які лінгвістичні групи, що беруть участь у тестуванні, можуть проходити його в різні дні, кожна мовна група повинна мати однакову кількість запропонованих тестових дат протягом року.

4.4. Докладно опишіть мовні адаптації та їх обґрунтування в керівництві з тестування, як це рекомендується видавцем тесту.

4.5. Адміністратор тесту повинен дотримуватися найкращих умов, пов'язаних з проведенням тесту[3] для всіх груп. Крім того, адміністратор тесту відповідає за наступні дії, пов'язані з тестуванням усіх мовних і культурних груп до самого тестування:

- Ті з аспектів фізичного середовища, які впливають на проведення тесту або на інструмент, повинні бути максимально подібними для різних цільових груп населення, як це зазначено в керівництві з проведення тестування. Якщо виникає така ситуація, що не прописана в інструкції з проведення тестування, адміністратор тесту повинен звернутися за порадою до користувача або до розробника тесту, або докласти зусиль для організації тестування з мінімальною можливістю виникнення будь-яких збоїв для осіб, що його проходять.
- Допомога у забезпеченні точного визначення інструкцією з проведення тесту всіх аспектів його адміністрування, що можуть вимагати уважної перевірки пристосувань, які можуть бути необхідними для членів нових мовних або культурних груп.
- Забезпечення належної підготовки тестових наглядачів, а також їх обізнаності та сприйнятливості до потреб учасників тестування з усіх мовних груп.
- Консультування учасників тестування з усіх мовних або діалектичних лінгвістичних груп, для яких тест вважається доцільним.

4.6. Керівництво з тестування має містити чіткі, вичерпні та зрозумілі інструкції для наглядачів тестування, щоб зменшити джерела можливих помилок та чітко означити права та обов'язки учасників тестування.

[3] Більш детальну інформацію див. в Рекомендаціях МТК щодо використання тестів: www.intestcom.org

4.6.1. Наглядачі тестування повинні чітко дотримуватися процедур, що містяться в керівництві з проведення тестування.

4.6.2. Коли це можливо, наглядачі тестування повинні зачитувати тестові інструкції основною мовою учасників тестування, щоб звести до мінімуму вплив небажаних джерел відмінностей серед різних груп населення, навіть якщо зміст тесту призначений для надання доказів наявності знань або навичок на не основній для учасників тестування мові. Наглядачі тестування повинні дослівно зачитувати будь-які тексти, інструкції або приклади, надані для проведення тестування, мовою тесту або мовою кожної з груп учасників тестування.

4.6.3. Щоб дотримуватись стандартизованих умови та уникнути джерел відмінностей, неприпустимо, щоб наглядач тестування визначав та використовував власні критерії для інструкції або пояснень, в тому числі підказки.

4.7. Наглядачі тестування повинні пояснити учасникам тестування їх права та обов'язки. Наглядачі мають бути чутливими до ряду факторів, пов'язаних зі стимульними матеріалами, процедурами проведення тесту та способом надання відповідей, що можуть вплинути на достовірність висновків, зроблених на основі балів для всіх груп населення.

Тестові пристосування

4.8. Усі пристосування для проходження тестування мають бути розроблені та задокументовані з метою забезпечення достовірного вимірювання цільового конструкту серед представників усіх мовних та культурних груп.

4.9. Перед проведенням будь-якого тесту визначте, які види пристосувань при проведенні тесту допустимі для учасників тестування з будь-якої культурної чи лінгвістичної групи, та не вплинуть на вимірюваний конструкт. Процедури, що стосуються способів реалізації цих пристосувань, мають бути визначені до початку тестування.

4.9.1. Пристосування можуть включати (але не обмежуються) використання словників (електронних), переклад лексиконів зі складними словами на різні мови, додатковий час, глосарій, надання інструкцій тестування рідною мовою (-ами) учасника тестування, проведення альтернативних форм тесту рідною мовою (-ами) учасника тестування, або використання допомоги перекладачів.

4.9.2. Якщо використовується допомога перекладачів, вони повинні вільно володіти рідною мовою учасника тестування і мовою тесту. В ідеалі, перекладач повинен мати досвід перекладу цих двох мов і розуміти ненавмисні наслідки, що можуть бути викликані поганим перекладом.

4.10. При проведенні тесту окремим учасникам необхідно заздалегідь розглянути умови надання пристосувань на індивідуальній основі, оскільки кожна людина може мати різний ступінь акультурації або рівень знання мови тестування.

4.11. На всіх етапах процесу тестування ставтесь до всіх учасників тестування, які будуть використовувати пристосування, справедливо й аналогічно іншим учасникам тестування.

Принцип 5: Інтерпретація оцінок та звітність

Інтерпретація та звітність оцінювання

5.1. Розробіть звіти з оцінювання та супроводжуючі матеріали для інтерпретації для різних мовних і культурних груп, що проходять тест, особливо у випадку, коли результати мають важливі наслідки для окремих учасників тестування.

5.2. Необхідно проводити огляд усіх існуючих звітів з оцінювання та матеріалів для інтерпретації з метою отримання та надання інформації про планування та розробку будь-яких нових матеріалів.

5.3. Розробники звітів з оцінювання мають бути знайомі з відмінностями в мовному статусі тих, хто буде отримувати звітні матеріали (наприклад, учасники тестування, їх батьки або місцеві органи влади). В процесі розробки звітів має враховуватись те, яким чином повідомити результати звітів про оцінювання та пов'язаних з ними матеріалів, щоб максимізувати розуміння та корисність цієї інформації для всіх одержувачів. Наприклад, для охоплення різних аудиторій, включаючи користувачів з обмеженим рівнем грамотності та лінгвістичні групи, мови яких не мають письмової форми, інформація може передаватися у різних формах, таких як усна, візуальна, а також письмова.

5.4. Використовуйте фокус-групи та інші методи збору даних (обговорення, інтерв'ю, спостереження тощо), щоб визначити незрозумілі елементи звітів з оцінювання (наприклад, технічні терміни, складні речення), які потенційно можуть викликати труднощі для будь-якої мовної групи, що є учасником тестування.

5.5. Збирайте докази інтерпретативності та використання звітних матеріалів для всіх мовних груп, коли такі матеріали використовуються, і надалі використовуйте ці дані для інформування ревізій результатів оцінювання та майбутньої практики звітності.

Розробка та зміст звіту

5.6. У звітних документах використовуйте мову, що прийнятна одержувачам звіту, щоб повідомити про тестові цілі та доцільне використання тесту. Особливо зверніть увагу на ті мовні чи культурні групи, які можуть бути гірше ознайомлені з тестовими шкалами, звітами та інтерпретаційними настановами.

5.7. Включайте у звіти про оцінювання інтерпретаційні настанови або інформацію для референтної групи населення та для різних мовних груп, особливо для найбільш поширених. Метою мають бути паралельні форми звітів з оцінювання та інтерпретаційні настанови написані різними мовами. Якщо можливо, перекладіть звіти з оцінювання на кожну мову та опишіть технічні терміни, для того, щоб перекладачі мали точне і правильне розуміння будь-якої технічної термінології, пов'язаної з оцінюванням.

5.8. Оформлюйте звіти таким чином, щоб головні результати були візуально помітними та зрозумілими для різних мовних груп.

5.9. Враховуйте цілі та використання тесту при відборі чисельної, графічної та текстової інформації, що буде включена в звіти з оцінювання. Обрані варіанти мають враховувати релевантність різних типів результатів для одержувачів, їх рівень володіння мовою та технічні міркування щодо ступеню деталізації даних.

5.9.1. Там, де використовуються графічні відображення результатів, чітко позначайте числові бали, бали по шкалах та інші елементи, використовуючи просту та зрозумілу мову, яка підходить для всіх мовних груп.

5.10. Уважно розгляньте групові порівняння між учасниками тестування на основі демографічних (географічних, мовних, расових / етнічних тощо) особливостей в контексті тестового змісту, щоб уникнути можливого неправильного тлумачення спостережуваних моделей поведінки. У звітних матеріалах, підготовлених для всіх мовних та культурних груп, повідомляйте результати в ясній, безоціночній формі та надавайте пояснення результатів, щоб уникнути неправильного тлумачення даних.

5.11. При звітуванні результатів, в тому числі по субшкалах, розробники тесту зобов'язані повідомляти про технічні застереження (наприклад, вища, ніж загальна оцінка, помилка вимірювання), що мають відношення до субшкал. Вони також повинні повідомляти, які інтерпретації субшкал доречні для всіх мовних груп.

Надання та доступ до звітів та інтерпретаційних матеріалів

5.12. Переклад/адаптація звітів та інтерпретаційних матеріалів на мови меншин, які представлені серед досліджуваних, повинні здійснюватися кваліфікованими перекладачами, при використанні найкращих засобів для забезпечення значущості балів та належних інтерпретацій.

5.13. Забезпечте звітами з оцінювання та допоміжними інтерпретаційними матеріалами тих, хто має їх отримати, на бажаній для них мові або надайте їм чітку інформацію про те, як отримати ці звіти.

5.14. При розробці допоміжних інтерпретаційних матеріалів для кожної представленої мови, розробник/видавець тесту повинні пересвідчитися в тому, що фактичний механізм доставки забезпечує доступ до цих матеріалів всім користувачам з різних мовних та культурних груп.

Використання тесту

5.15. Надайте вказівки щодо значення та використання отриманих балів зрозумілою та однозначною мовою, що відображає рівні володіння мовою цільових користувачів.

5.16. Обирайте тест з огляду на його відповідність цілям тестування, беручи до уваги сам тест і фонові характеристики цільового населення, включаючи всі мовні групи.

5.16.1. Якщо тест використовує нормативні показники, слід провести перевірку нормативної вибірки, щоб переконатися, що вона є репрезентативною для цільового населення, включаючи всі мовні групи.

5.17. За використання тесту відповідають розробник та користувач. Обидва з них повинні надати достатньо доказів мети використання тесту відповідно до його розробки та конструкції.

5.17.1. Розробник оцінювання відповідає за те, щоб надати чітку інформацію про передбачене використання тесту та інтерпретацію його балів. Він також має надати докази, що підтверджують заявлені характеристики конструкту для різних груп населення, які проходять тест.

5.17.1.1. Якщо інтерпретація результатів оцінювання не сходиться з конструктом, для вимірювання якого призначалося тестування, особа, яка проводить тестування, відповідає за рішуче застереження зацікавлених сторін щодо відсутності доказів для підтримки тверджень, що можуть бути винесені стосовно мовних або культурних груп.

5.17.1.2. Особа, яка проводить тестування, має право скасувати або визнати недійсними оцінки в ситуації, коли тестування використовується для несанкціонованого використання оцінювання або його балів.

5.17.2. Особа, яка проводить тестування, несе відповідальність за документоване обґрунтування вибору конкретного типу оцінювання, в тому числі за надання доказів, що підтверджують твердження про конструкт, які можуть бути зроблені щодо цього оцінювання.

5.17.3. Якщо відхилення від зазначеної мети тесту є бажаним, особа, яка проводить тестування, має надати обґрунтування та докази на підтримку нового способу використання оцінювання, включаючи пояснення щодо інтерпретації отриманих балів для кожної мовної групи. В такій ситуації має бути проведене дослідження валідності, щоб переконатися в тому, що оцінювання відповідає передбачуваній меті використання.

5.17.3.1. Якщо існують докази недоречності нового способу використання оцінювання, будь-яка з зацікавлених сторін (включаючи учасників тестування, осіб, що проводять чи використовують тестування) повинна повідомити особі, що використовує тест та контролює чи переглядає прийняті на основі його результатів рішення, про неналежне використання тесту.

5.18. В ситуації, коли особи, які беруть участь у тестуванні, відносяться до культурно чи лінгвістично різноманітних груп, розробники та видавці оцінювання, якщо це можливо, мають надавати чітку інформацію щодо належного та неналежного використання тесту та інтерпретації балів.

5.19. Розробник тесту повинен надати відповідні технічні настанови щодо видів адаптацій, які були внесені до оцінювання, інструкції щодо інтерпретації балів, дані про те, для кого призначене адаптоване оцінювання, а також інформацію про те, наскільки дійсними та валідними можуть бути висновки, зроблені з отриманих оцінок.

5.20. Розробник тестування та особа, яка використовує тест повинні прагнути до чіткого розуміння балів та результатів тестування, їхньої достовірності та впливу, який ці оцінки матимуть на учасників тестування з будь-якої мовної та культурної груп.

5.20.1. Особа, яка використовує тест, несе відповідальність за контроль та гарантування того, що неправильні тлумачення або неналежне використання тесту не матимуть місця під її керівництвом.

5.20.2. Особа, яка проводить тестування, повинна забезпечити надання громадськості відповідних пояснювальних матеріалів до результатів випробувань, що публікуються в загальному доступі, щоб уникнути неправильного тлумачення, особливо якщо є відмінності в результатах тестування осіб з різних мовних груп.

5.21. Якщо особа, яка проводить тестування, робить запит на будь-які серйозні адаптації, що стосуються формату, мови або способу проведення тесту, вона має надати вагоме обґрунтування цього запиту. Крім того, особа, що проводить тест, повинна провести перевірку достовірності та надійності модифікованого варіанту оцінювання.

Принцип 6: Контроль якості для досягнення порівнянності та справедливості в оцінці тестів

6.1. Для забезпечення стандартизованих умов тестування, розгляньте варіант створення контрольного переліку, який допоможе забезпечити використання належних критеріїв для різних мовних та культурних груп при всіх процесах оцінювання.

6.2. Для оцінок, що надаються оцінювачами, зважте можливість розгляду сирих балів, щоб визначити, чи існують взаємодії між оцінювачами та відповідями певних лінгвістичних груп. Якщо виявляються суттєві та систематичні відмінності в оцінках, слід провести дослідження, щоб забезпечити, що відмінності у оцінках не будуть мати ненавмисних негативних наслідків для учасників тестування з цих мовних груп.

6.3. Контроль якості оцінок за шкалою має проводитися до остаточної звітності оцінок для груп L1 та L2.

6.3.1. Технічний посібник до тесту має містити пояснення щодо методів, які використовувались для визначення шкал та їх еквівалентності сирим балам, і, якщо це необхідно, форму, яку користувач тесту може використати для адаптації шкали до іншого контексту.

6.4. Коли це можливо, розглядайте оцінки кожної мовної групи окремо, а потім порівнюйте їх між собою, а також з референтною популяцією.

6.4.1. Де це можливо, порівняйте очікувані та спостережувані оцінки для кожної лінгвістичної групи один з одним, щоб знайти тенденції (на основі балів, отриманих при попередньому та поточному тестуванні).

6.4.1.1. Перевіряйте точність, надійність оцінок та швидкість проходження тестування для всіх груп та проведень тесту.

6.4.1.2. Відмінна швидкість проходження тестування між учасниками тестування L1 та L2 може свідчити про те, що оцінки можуть бути не порівнянними.

6.4.2. Оцініть зміни у балах для учасників тестування з будь-якої лінгвістичної групи, а також зміни у загальній групі.

6.5. Документуйте всі статистичні дані та аналізи для майбутніх досліджень та перевірок контролю якості.

6.6. Перевірте засоби та стандартне відхилення змін у балах осіб, які проходять тестування більше одного разу.

6.6.1. Стільки, скільки оцінювання продовжує проводитись та використовуватись, збирайте поточні докази про прохідні/розрізнюючі (selection) бали для учасників тестування з різних мовних груп.

6.7. Проводьте дослідження з контролю якості окремо для кожного етапу процесу оцінювання, щоб вони служили основою для забезпечення порівнянності балів.

6.7.1. Якщо тест буде проводитись всім мовним популяціям однією мовою (наприклад, референтною мовою), проводьте дослідження, щоб забезпечити порівнянність балів між групами.

Принцип 7: Підготовка до тестування

7.1. Щоб ознайомити всі лінгвістичні групи з питаннями та форматом проведення тесту, особа, яка проводить тестування, повинна забезпечити учасників тестування з усіх лінгвістичних та культурних груп затвердженими практичними завданнями, зразками або іншими матеріалами для підготовки до тесту відповідно до рекомендованої практики перед проходженням самого тесту. Всі матеріали (в тому числі тестові інструкції) мають бути адаптовані таким чином, щоб не змінювати оцінюваний конструкт. Окрім того, мають бути надані описи матеріалів, що відображають прийнятні пристосування, якщо це необхідно.

7.2. Видавець тесту повинен надати повний опис тесту, його характеристик та призначення. Це особливо важливо при оцінюванні осіб з різних мовних або культурних груп, оскільки це дає можливість мінімізувати потенційні розбіжності, що виникають внаслідок відмінності у розумінні типів або формату питань, які можуть бути незнайомими для деяких досліджуваних груп населення.

7.2.1. Матеріали для підготовки до тестування повинні містити інформацію про специфіку тесту. Це включає в себе інформацію про час, кількість розділів у тесті, зразки кожного з типів питань, які будуть в кожній з частин тесту, і інформацію про те, як і коли будуть повідомлені результати. Це дозволить усім групам ознайомитися зі змістом і форматом тесту, способом його проведення, його довжиною, інструкціями до кожного типу питань, тривалістю та оцінкою тестування (в тому числі і з інформацією про віднімання балів, якщо це застосовується у тесті).

7.2.1.1. Якщо це можливо, підготуйте список стратегій тренування з проходження тесту для таких учасників тестування, які з огляду на їх мовний досвід, можуть бути незнайомі з підготовкою до певних видів тестів. Це можуть бути пропозиції про створення підготовчих груп (що складаються з учасників тестування з тієї ж самої, з інших або зі змішаних мовних груп) та використання допоміжних навчальних ресурсів.

7.3. Якщо тест буде проводитись з використанням комп'ютеру, вкажіть, яким він буде – адаптивним або лінійним (фіксованим), і поясніть, що це означає на практиці. Для адаптивних тестів поясніть, що складність тесту варіюється в залежності від того, наскільки добре учасник тестування відповідає на попередні питання, і що рівень складності наступних питань може швидко збільшитися, щоб намагатися відповідати рівню кваліфікації учасника тестування.

7.4. Там, де це можливо, надайте правильні відповіді та обґрунтування правильних відповідей для кожного питання тестового зразка, щоб забезпечити ознайомлення з тестом та його питаннями учасників тестування з усіх мовних груп.

7.5. Чітко роз'ясніть відмінності між підходами підготовки до тесту (наприклад, між коучингом та навчанням) та визначте той з них, який вважається прийнятним для тесту.

Словник термінів та визначень

L1 – посилання на рідну / першу мову (-и) учасників тестування.

L2 – посилання на другу або нерідну мови учасника тестування.

Read-aheads (розсіяні відповіді) – використовується як перевірка якості для забезпечення налаштованості оцінювачів. Це досягається завдяки тому, що оцінювачі виставляють бали есе, які до цього були оцінені керівником з оцінювання, щоб переконатися, що їхні бали не дрейфують (точність їх оцінок не змінюється).

Read-behinds – повторне виставлення балів, оцінювання есе, що проводиться експертами на основі випадкової вибірки відповідей, які вже були оцінені іншими члени команди оцінювачів.

Адаптації (адаптовані тести) – будь-які зміни, що вносяться до проектування оцінювання стосовно змісту, формату чи особливостей проведення тесту, з метою полегшити, збільшити доступ до матеріалів оцінювання для культурних або мовних груп, які можуть відрізнятися від основного населення. Наслідки модифікації самого тесту або його проведення впливають на інтерпретацію результатів. Ці наслідки повинні спільно розглядатися розробником тесту та особою (-ами), що його проводять.

Адаптивний тест – тест, що автоматично пристосовується та зазвичай керується комп'ютером. Це такий тест, в якому алгоритм визначає, яким буде кожне наступне тестове питання, що пропонується досліджуваній особі (менш складне, складніше або питання такого ж рівню складності), на основі відповідей на попередні питання. Мета цих тестів полягає в тому, щоб забезпечити кращу оцінку здібностей кожного окремого учасника тестування.

Аналітичне оцінювання (оцінка за характеристиками) – метод оцінювання есе, в якому кожна характеристика (наприклад, граматика, якість аргументації, стиль письма тощо) оцінюється та підраховується окремо, а отримані бали об'єднуються для загальної оцінки. Цей метод є протилежним цілісній оцінці.

Багатоаспектне вимірювання Раша (MFRM) – підхід, що дозволяє вивчати такі ефекти оцінювачів як жорсткість/поблажливість, ефект ореолу та центральна тенденція, а також досліджувати корисність рейтингових шкал і наявність відмінностей у роботі оцінювачів, наприклад, коли деякі оцінювачі мають упереджене ставлення до певних груп учасників тестування, зокрема, до культурних або мовних груп.

Багатоступеневе адаптивне тестування – подібне до адаптивного тестування, алгоритм визначає, які групи тестових питань пропонуються учасникам тестування, на основі результатів попередніх питань, які були менш складними, більш складними або мали такий же самий рівень складності.

Багаточленні питання – тестові питання, які мають більше двох категорій відповідей або оцінок.

Бали – інше слово для позначення рейтингів, оцінок або результатів.

Валідність – валідність оцінки відноситься до того, наскільки теорія та емпіричні дані підтверджують передбачене значення та використання результатів оцінювання.

Вихідні (оригінальні) та цільові мови та культури – вихідна мова – це мова оригіналу тесту. Цільова мова – це мова, на яку тест перекладається.

Вказівки – посібник, що містить критерії оцінки есе або інших відкритих питань тестування.

Диференційний аналіз функціонування питань (DIF) – коли група випробуваних, що мають однакові здібності, обирають правильний варіант відповіді з різною частотою, порівняно з еталонною групою учасників тестування. Ці відмінності можуть з'являтися з джерел, що мають чи не мають відношення до конструкту. У широкомасштабних оцінюваннях зазвичай використовуються два методи ідентифікації однорідного DIF: (1) **метод Mantel-Haenszel** - у великих фокусних і контрольних групах або (2) **стандартизація** - коли одна або обидві групи малі. Докладніше див. Dorans and Holland (1992) або Osterlind і Everson (2009). Для дослідження як однорідного, так і неоднорідного DIF в рамках підтверджуючого факторного аналізу (CFA), переважним є використання тесту на коефіцієнт вірогідності або тесту Wald на відмінності розрізнення та параметри складності.

Діалект – мова, яка використовується в конкретному регіоні або в окремій соціальній групі.

Доленосні оцінювання (на противагу рішень, що мають меншу вагу) – доленосні оцінювання мають наслідки для важливих рішень щодо учасника тестування як-то рішення про допущення чи недопущення до чогось, отримання рангу, стипендії або виставлення діагнозу. Оцінювання, що мають меншу важливість призначені для оцінки індивідуума в певний момент часу для визначення майбутніх рішень, що будуть зроблені вчителем, професором або психологом (наприклад, зворотній зв'язок з учнями, курсові завдання).

Еталонні відповіді – відомі також як якірні відповіді. Включають попередньо відібрані есе, які використовуються як приклади виставлення різних балів за вказівками до оцінювання, та застосовуються для підготовки та налаштування оцінювачів.

Ефект ореолу – тип упередження, при якому оцінювач надає подібні рейтинги концептуально різним критеріям або аспектам виконання тесту (наприклад, загальне враження оцінювача щодо учасника тестування або його відповідей впливає на аналогічне оцінювання усіх інших його показників).

Ефект оцінювача – ситуація, коли оцінювач систематично присвоює есе бали, що не обов'язково є об'єктивним відображенням критеріїв оцінювання. Типи упереджень оцінювача включають центральну тенденцію, ефект ореолу, ефекти поблажливості та жорсткості.

Керівництво з проведення тесту – містить правила та процедури для адміністраторів тестів, наглядачів та інших людей, як будуть проводити тест. Керівництво має надавати інструкції для забезпечення стандартизованих умов тестування за такими параметрами як

користування тестовими матеріалами, час проведення тесту, обстановка тестування, наглядання, надання учасникам пристосувань та проведення процедур щодо порушень та підозрюваного шахрайства.

Конструкт – знання, навички, здібності або властивості, на вимірювання яких спрямований тест. Конструкти не спостерігаються безпосередньо і є прихованими.

Коучинг – короткострокові підходи до підготовки до тестів, такі як легкі стратегії проходження тестування або швидкі виправлення (fixes), що допомагають підвищити бали.

Лінійний тест – у лінійному тесті всі питання тесту подаються всім учасникам в однаковому порядку, незалежно від результатів відповіді на попередні тестові питання.

Межі діапазону – перед оцінюванням добираються есе в якості еталонних відповідей, а потім використовуються для навчання та налаштування осіб, які будуть присвоювати бали кожному есе.

Межовий бал – це попередньо обрана точка (або точки) на шкалі оцінки, яка використовується для розрізнення класів учасників тестування. Межовий бал може використовуватись для класифікації учасника випробування як такого, що володіє певними характеристиками на основі його результатів тестування, наприклад, демонструє мінімальну компетентність у знаннях або здатностях, що відносяться до певного конструкту. При оцінюванні можуть бути використані декілька межових балів, щоб класифікувати учасників тесту за різними заздалегідь визначеними категоріями на основі групових стандартів.

Модульна конструкція тесту – тест, що складається з декількох секцій, всі або деякі з яких можуть бути замінені іншими еквівалентними секціями тестування.

Навчання – довгостроковий підхід до підготовки до тесту, який має на меті покращити знання та навички.

Наглядач тесту – особа, відповідальна за проведення оцінювання для учасників тестування та забезпечення відповідного виконання всіх процедур адміністрування тесту. Наглядачі тесту також несуть відповідальність за те, щоб відповідати на будь-які процедурні питання, які можуть з’явитися у учасників тестування. Вони також слідкують за тим, щоб учасники тестування виконували свої власні завдання і не копіювали відповіді інших учасників.

Надійність – фундаментальне поняття, що відноситься до точності вимірювання тесту. Існують різні методи оцінки надійності тестування (наприклад, надійність між оцінювачами, надійність повторного тестування, надійність паралельних форм). *Надійність між оцінювачами* вказує на згоду щодо балів (узгодженість) між різними оцінювачами стосовно відповідей учасників тестування на відкриті питання. *Надійність повторного тестування* свідчить про постійність балів тесту, що проводиться багаторазово. *Надійність паралельних форм* вказує на узгодженість балів між різними тестовими формами одного оцінювання.

Надійність між оцінювачами – відноситься до узгодженості балів, наданих різними оцінювачами, або згоди щодо балів між двома або більше оцінювачами.

Невідповідність балів – відмінність у оцінках, що присвоюються оцінювачами двом учасникам, на певну кількість балів (різниця більше одного балу).

Нормований тест (нормативне оцінювання) – тест, стандартизований показник якого свідчить про те, наскільки добре учасники тестування справляються з питаннями, порівняно з результатами статистично вибраної групи учасників тестування з групи, що має нормальний розподіл.

Об'єктивність – невід'ємна якість самого об'єкта, не пов'язана з будь-яким спекулятивним підходом; ця ознака працює на користь рівності та справедливості тестування. Серед її властивостей є: (а) відсутність упередженості в інтерпретаціях та прийнятті рішень, (б) зосередження на неупередженості оцінювачів, та відсутності у них власних припущень і цінностей; (в) розрізнення двох протилежних або навіть суперечливих ідей або теорій на основі точного визначення об'єкта (Gaukroger, 2012). Об'єктивність – це атрибут, використання якого не обмежується запитаннями з множинним вибором або іншими закритими формами питань.

Користувач тесту – фахівці з організацій, які обирають інструмент для вимірювання конкретних ознак з певною метою. Вони несуть відповідальність за правильне тлумачення результатів тестування та за висновки про те, що ці оцінки відображають. Вони також несуть відповідальність за те, щоб оцінювання проводилось належним чином у стандартизованих умовах.

Офіційна мова – мова або одна з мов, що затверджені урядом країни для використання в юридичних та офіційних документах, якою викладають в школах та яку використовуються в правовій системі.

Оцінювання за конкретними завданнями – стосується оцінювання виконання завдання за окремою ознакою або набором ознак, що є особливо актуальними для цього завдання. Це може бути реалізовано за допомогою використання цілісної шкали (оцінка первинної ознаки) або за допомогою конкретного набору завдань аналітичної шкали (оцінка множинних ознак).

Оцінювачі – особи, які оцінюють есе та інші завдання. На них також посилаються як на суддів, читачів, маркерів, грейдерів, скрорерів, чекерсів.

Переклад – такий процес переведення оцінювання з однієї мови на іншу мову (або мови), при якому вимірюваний конструкт залишається незмінним, складність кожного перекладеного питання є такою, як і в оригіналі, а оцінки обох тестів є порівнянними, тобто бали, отримані після проходження обох тестів, можна інтерпретувати однаково.

Поблажливість (жорсткість) – загальна тенденція оцінювача оцінювати есе або занадто жорстко (суворо) або занадто розслаблено (поблажливо).

Порівнянність – порівнянність балів вказує на те, наскільки подібними можуть бути висновки зроблені на основі балів у аналогічних оцінюваннях. Інтерпретація балів оцінювань адаптованих для лінгвістично/культурно різноманітних груп має нести те ж саме значення, що і інтерпретація в оригінальному оцінюванні, з яких ці бали були отримані, оскільки обидва оцінювання мають досліджувати один і той же конструкт.

Пристосування – адаптації, що вносяться до проектування оцінювання або його проведення, та не змінюють вимірюваного конструкту або інтерпретації балів цього оцінювання.

Рейтингова шкала – весь діапазон можливих балів для призначеного питання, таких як, наприклад, есе. Іноді супроводжується описом характеристик виконання, або поведінки, що оцінюється деякими або всіма балами.

Сирі бали – загальна кількість питань, на які була надана правильна відповідь під час тестування або сума балів за питання, коли використовуються багаточленні питання. Можливі варіації способів виведення сирих балів тестування з балів за питання (наприклад, може використовуватися середнє значення); однак в подальшому сирі бали будь-яким чином не коригуються та не трансформуються.

Соціокультурний – використовується для опису поєднання соціальних і культурних факторів.

Соціолінгвістика – вивчення того, як на використання мови впливає широкий спектр соціальних ситуацій, включаючи відмінності між групами за регіонами, соціальними класами, статтю та роду занять.

Справедливість – поняття справедливості в оцінюванні має на увазі те, що висновки, зроблені на основі результатів оцінювання, є однаковими, незалежно від досвіду учасника тесту або його приналежності до певної групи. По відношенню до культурної та лінгвістичної різноманітності, це означає, що якщо особа проходить адаптоване тестування, то результати будуть належним чином демонструвати її знання або здатності, що відносяться до цільового конструкту.

Стандартизовані бали – перетворення сирих балів учасників тестування на стандартизований діапазон балів (шкала). Це дає змогу змістовно порівняти всіх учасників тестування з нормативним населенням а також результати між різними тестовими формами того ж самого оцінювання.

Субшкала – оцінка конкретного конструкту, що є частиною загальної складеної оцінки тесту.

Теорія узагальнення (G-теорія) – дозволяє розпізнавати множинні джерела похибки вимірювання, оцінюючи величину кожного з них окремо, і надаючи дані для мінімізації похибки вимірювання при оцінюванні.

Тести орієнтовані на критерій – тест, стандартизовані оцінки якого базуються на заздалегідь визначеному наборі критеріїв (або стандартів) проходження тесту, а не вираховуються порівняно з результатами інших учасників тестування.

Узгодженість – ступінь, до якої найголовніші характеристики оцінювання (наприклад, питання, оцінювачі, час тестування) є порівняними в умовах тестування.

Упередженість інструменту – виявлення відмінностей в результатах тестування населення при використанні оригінального тесту та його адаптованої форми, що обумовлені непов'язаними з конструктом факторами. Це може бути обумовлено різним рівнем знайомства зі стимульними матеріалами або процедурами надання відповідей, відмінними стилями надання відповідей (наприклад, набором відповідей, притаманних культурі, як-то різний рівень бажання та готовності до саморозкриття) або різним рівнем соціальної бажаності.

Упередженість оцінювача – ситуація, коли бали, що присвоюються відповідям учасників тестування оцінювачем, змінюються через певний аспект ситуації оцінювання, що не має відношення до вимірюваного конструкту; наприклад, коли оцінювачі призначають більш суворі оцінки членам лише деяких груп.

Учасники тестування – інше слово для позначення випробуваних, кандидатів або респондентів.

Центральна тенденція – уникнення оцінювачами крайніх оцінок рейтингової шкали та надання переваги оцінкам поблизу середини шкали.

Цілісна оцінка – оцінювання, при якому есе виставляється єдина, загальна оцінка, що відображає всі аспекти есе, відповідно до опису цієї оцінки у відповідних вказівках.

Читачі – див. оцінювачі.

Швидкісне тестування – оцінювання, в якому результат учасника тестування залежить не тільки від правильності відповідей, але й від швидкості їх надання. З точки зору справедливості, швидкісне тестування може поставити у невигідне положення учасників тестування L2, особливо, коли володіння мовою не є цільовим конструктом тесту.

Посилання

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*. Washington D.C.

Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and Standardization. *ETS Research Report Series*, *1992*(1), i-40.

Eckes, T. (2015). Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments (2nd ed.). Frankfurt am Main, Germany: Peter Lang.

Educational Testing Service (2015). *ETS Guidelines for Fair Tests and Communications*. Retrieved from:
https://www.ets.org/s/about/pdf/ets_guidelines_for_fair_tests_and_communications.pdf

Educational Testing Service (2009). *ETS International Principles for Fairness Review of Assessments - A Manual for Developing Locally Appropriate Fairness Review Guidelines in Various Countries*. Retrieved from: https://www.ets.org/s/about/pdf/fairness_review_international.pdf

Educational Testing Service (2009). *Guidelines for the Assessment of English Language Learners*. Retrieved from: http://www.ets.org/s/about/pdf/ell_guidelines.pdf

Elosua, P. (2016). Minority language revitalization and educational assessment: Do language-related factors impact performance? *Journal of Sociolinguistics*. 20(2), 212-228.

Engelhard, G., & Wind, S. A. (2018). Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments. New York, NY: Routledge.

Gaukroger, S. (2012). *Objectivity. A very short introduction*. Oxford: Oxford University Press.
Haugen, E. (1966). Dialect, Language, Nation. *American Anthropologist*, 68(4), 922–935. Retrieved from http://www.jstor.org/stable/670407

International Test Commission (2017). *Guidelines for Translating and Adapting Tests, 2nd edition*. Retrieved from: https://www.intestcom.org

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). Assessing performance: Designing, scoring, and validating performance tasks. New York, NY: Guilford.

Luykz, A., Lee, O., Mahotiere, M., Lester, B., Hart, J., & Deaktor, R. (2007). Cultural and home language influences in children's responses to science assessments. *Teachers College Record*, 109(4), 897-926.

Oakland, T. (2016). Testing and assessment of immigrants and second-language learners. In: Leong, F. et al. (Eds.). *The ITC International Handbook of Testing and Assessment*. Oxford University Press.

Oliveri, M. E., Lawless, R. R., & Mislevy, R. J. (2018*). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments*. (Manuscript in press).
Oliveri, M.E., Lawless, R., & Young, J. (2015). *A validity framework for the use and development of exported assessments*. Princeton: ETS Office of Professional Standards Series. Retrieved from: https://www.ets.org/s/about/pdf/exported_assessments.pdf

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Los Angeles, CA: Sage.
Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing*, 3, 129-150.

Survey Research Center. (2016). *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved June 21, 2018 from http://www.ccsg.isr.umich.edu/.