Міжнародна тестова комісія

Міжнародні тестові домовленості

Керівництво Міжнародної тестової комісії з перекладу та адаптації тестів. Друга редакція

Переклад:

Кірієнко Пелагея

Варто цитувати як:
INTERNATIONAL TEST COMMISSION (2010). ITC GUIDELINES FOR TRANSLATING ANG ADAPTING TESTS (SECOND EDITION).
[HTTP://WWW.INTESTCOM.ORG]

ПОДЯКА

Рада Міжнародної тестової комісії хоче подякувати комітету з шести осіб, що працював над другою редакцією Керівництва з перекладу та адаптації тестів протягом декількох років: Девіду Бартраму, SHL, Великобританія; Гіраю Берберглу, Близькосхідний технічний університет, Туреччина; Жаку Грегуару, Левенський католицький університет, Бельгія; Рональду Хамблтону, Голові комітету, Массачусетський університет в Амхерсті, США; Хосе Муньїсу, Ов'єдський університет, Іспанія; а також Фонсу ван де Війверу, Університет Тілбурга, Нідерланди.

Крім того, подяка виноситься Чаду Букендалю (США); Енн Герман та її колегам з OPP Ltd. (Велика Британія); а також Ейпріл Зеніскі з Массачусетського університету (США), за ретельний та уважний розгляд попереднього варіанту документа. МТК вдячна усім іншим рецензентам, хто будь-яким чином зробив свій внесок у другу редакцію Керівництва МТК з перекладу та адаптації тестів.

АНОТАЦІЯ

Друга редакція Керівництва МТК з перекладу та адаптації тестів розроблялася у період з 2005 по 2015 рік. Метою було покращити та доповнити першу редакцію, враховуючи значний розвиток методів та технології тестування. Вісімнадцять вказівок було впорядковано та поділено на шість категорій для більшої зручності у використанні: передумова (3), розробка тесту (5), затвердження (4), проведення тестування (2), підрахунок балів та тлумачення (2), документування (2). Також надано чек-лист для більш ефективного втілення принципів, викладених у Керівництві.

ЗМІСТ

ПЕРЕДІСТОРІЯ

В останні 25 років сфера методики перекладу та адаптації тестів стрімко розвивалась завдяки публікації ряду книг, проведенню нових наукових досліджень та блискучим прикладам адаптації тестів (див., наприклад, van de Vijver & Leung, 1997, 2000; Hambleton, Merenda, & Spielberger, 2005; Grégoire & Hambleton, 2009; Rios & Sireci, 2014). Такий прогрес у цій сфері мав велике значення адже все більше зростала зацікавленість у (1) психології міжкультурних відмінностей, (2) великомасштабних міжнародних порівняльних дослідженнях (як, наприклад, TIMSS та OECD/PISA), (3) акредитаційних тестах, що використовуються у всьому світі (наприклад, у сфері інформаційних технологій такими компаніями як Microsoft та Cisco), та (4) справедливості умов тестування, досягнутої через дозвіл на обрання мови проведення оцінки для екзаменованих (напр., можливість для абітурієнтів Ізраїлю здавати екзамени для прийому в університет однією з шести мов).

У сферах якісного та кількісного підходів до дослідження інструментальних та методичних похибок адаптованих тестів та опитувальників відбувся значний технічний прогрес – почали використовуватися такі складні статистичні процедури як теорія відповідей на тестові завдання (IRT), моделювання структурними рівняннями, та теорія узагальнення (див. Hambleton et al., 2005; Byrne, 2008). Завдяки OECD/PISA були створені нові методики перекладу (див. Grisay, 2003); було запропоновано ряд заходів для завершення проектів з адаптації тестів (див., наприклад, Hambleton & Patsula, 1999; доступні взірцеві проекти, що можуть слугувати орієнтиром для практики адаптування тестів – наприклад, проекти OECD/PISA та TIMSS); окрім цього було багато інших значних досягнень.

Розробка першої редакції Керівництва (див. van de Vijver & Hambleton, 1996; Hambleton, 2005) почалася з порівняльної перспективи, що і є метою адаптації тесту – дозвіл та сприяння проведенню порівнянь між групами респондентів. Негласний шаблон проекту, для котрого і призначалося керівництво, послідовно розвивався як інструмент для застосування у контексті порівняння (існуючий інструмент має бути адаптовано під використання у новому культурному контексті). Однак стає все більш зрозумілим те, що у адаптації тестів значно ширша сфера застосування. Дуже важливим є те, що існуючі чи нові інструменти використовуються у багатокультурному суспільстві: при психологічній практиці з клієнтами, що відносяться до різноманітних етнічних груп, при проведенні навчальних оцінювань в групах студентів з цілком різним культурним походженням та диференціальним ступенем володіння мовою тестування, а також у сфері найму, що зараз як ніколи орієнтована на міжнародну співпрацю. Дані зміни в сфері застосування тягнуть за собою подальші зміни у розробці, проведенні тестування, валідизації та документуванні. Наприклад, можливими наслідками може бути адаптування завдань існуючого тесту (наприклад, через спрощення лексики/граматики) для покращення розуміння його змісту учасниками, що не є носіями мови оригіналу. Іншим важливим додатком до керівництва буде впровадження ідеї про одночасну розробку (тобто об'єднану розробку різномовних версій тесту). Повномасштабні міжнародні проекти все більше використовують даний метод для уникнення проблеми складнощів перекладу, що може виникнути при спробі перекласти/адаптувати вже готовий інструмент, написаний однією мовою.

Перша редакція Керівництва МТК з перекладу та адаптації тестів була опублікована ван де Війвером та Хамблтоном (van de Vijver & Hambleton, 1996), Хамблтоном (2002), та Хамблтоном, Мерендою та Шпільбергером у 2005. У період між 1996 та 2005 були внесені

лише невеликі редакційні зміни. Між тим, з 1996 р. було досягнуто значного прогресу. Насамперед було зроблено ряд корисних рецензій на Керівництво МТК. Серед них – праці Джинрі та Бертранда (Jeanrie & Bertrand, 1999), Танцера та Сіма (Tanzer & Sim, 1999) та Хамблтона (2002). Кожен з авторів відмітив велику цінність керівництва, але також запропонував ряд пропозицій щодо його вдосконалення. Хамблтон, Меренда та Шпільбергер (2005) опублікували головні матеріали міжнародної конференції МТК, що була проведена у Джорджтаунському університеті в США в 1999 р. Деякі з авторів розділів (напр., Cook & Schmitt-Cascallar, 2006; Sireci, 2005) сформували нові парадигми адаптації тестів та запропонували нову методологію. У 2006 році МТК провела міжнародну конференцію у Брюсселі (Бельгія), щоби зосередити загальну увагу на Керівництві МТК з перекладу та адаптації тестів. Понад 400 учасників конференції з більш ніж 40 країн зосередились на темі адаптування тестів, в результаті чого було запропоновано багато нових методологічних ідей, вказівок та прикладів успішного їх застосування. На міжнародних симпозіумах, проведених у період між 1996 та 2009 роками, було представлено чимало статей, наукових доповідей та інших матеріалів (див., наприклад, Grégoire & Hambleton, 2009), а також первісну версію другої редакції Керівництва МТК іспанською мовою (Muniz, Elosua, and Hambleton, 2013).

У 2007 р. Рада МТК зібрала комітет з шести персон, обов'язком якого було доповнити Керівництво МТК оновленою базою знань та розповсюдити розроблені дослідниками нові методи та прийоми, що сформували передовий досвід у сфері. Серед інновацій: (1) моделювання структурними рівняннями для виявлення факторної еквівалентності поміж різномовних груп, (2) розширений підхід до встановлення рівню диференціального функціонування завдань за допомогою політомної моделі для рейтингової шкали в різномовних групах, а також (3) нові адаптаційні техніки, створені під егідою таких міжнародних проектів з оцінювання як OECD/PISA та TIMSS. Комітет також представив презентації та чорнові версії нових керівних принципів на міжнародних засіданнях психологів в Празі (2008) та Осло (2009), отримавши там важливі відгуки та оцінки від колег.

Розділ принципів з проведення тесту зберігся в другій редакції, але частково співпадаючі за своєю суттю вказівки були суміщені, за рахунок чого загальна кількість знизилась з шести до двох. «Документування/тлумачення оцінок» було останнім розділом в першій редакції. У другій редакції ми розділили його на один, що повністю зосереджений на темі підрахунку балів та тлумаченні, та другий, в якому йдеться мова про документацію. До того ж, два з чотирьох оригінальних принципів в цьому розділі були ґрунтовно змінені.

Як і в першій редакції, ми хочемо розмежувати поняття перекладу та адаптації тесту для наших читачів. Переклад тесту, можливо, є більш поширеним терміном, у той час як адаптація тесту є ширшим поняттям, що охоплює не просто переклад тесту з однієї мови на іншу, але також урахування культурних реалій країн-носіїв оригінальної та цільової мов. Адаптація тесту включає в себе наступні аспекти: визначення того, чи може адаптований під нову мову та культуру тест вимірювати той самий фактор, що і в оригінальній версії; відбір перекладачів; обрання способу оцінки результату роботи перекладачів тесту (наприклад, метод зворотного перекладу); створення будь-яких необхідних умов для пристосування до потреб певних груп людей; коригування формату тесту; здійснення перекладу; перевірку еквівалентності тесту на другій мові, адаптованого до відповідної культури, та проведення інших необхідних досліджень валідності. Переклад ж тесту має більш вузьке значення, що зводиться до вибору мови, якою буде перекладено оригінальний тест зі збереженням лінгвістичного сенсу.

Переклад тесту є лише частиною його адаптації, але сам по собі він може бути досить спрощеним підходом до відтворення змісту оригінального тесту новою мовою без огляду на освітню чи психологічну еквівалентність.

КЕРІВНИЦТВО

Вступ

В нашій праці дане керівництво подається як практика, необхідна для проведення та оцінки адаптації (що також іноді називають «локалізацією») або одночасної розробки психологічних та навчальних тестів для використання різними групами населення. Вісімнадцять вказівок було впорядковано та поділено на шість широких тем: передумова (3), розробка тесту (5), затвердження (емпіричний аналіз) (4), проведення тестування (2), підрахунок балів та тлумачення (2), документування (2).

У першому розділі під назвою «Принципи передумови» підкреслюється, що будь-які рішення мають бути прийняті ще до початку самого процесу перекладу чи адаптації. Другий розділ, «Принципи розробки тесту», фокусується на дійсному процесі адаптування тесту. Третій розділ «Принципи затвердження» включає в себе принципи узагальнення емпіричних доказів еквівалентності, надійності та валідності тесту в різноманітних мовах та культурах. Назви останніх трьох розділів говорять самі за себе: «Проведення тесту», «Підрахунок балів та тлумачення» та «Документування». Теми документування в сфері адаптації психологічних та навчальних тестів завжди уникали найбільше, тому ми би хотіли, щоби редактори журналів та фінансуючі установи вимагали більше від документування процесу адаптації тесту.

До кожної вказівки надані пояснення та підказки для ефективнішого втілення їх на практиці.

Принципи передумови

PC-1 (1) Перед здійсненням адаптації отримайте необхідний дозвіл від носія прав інтелектуальної власності на тест.

Пояснення. Права інтелектуальної власності – це комплекс прав людини або групи людей на власні твори, винаходи, чи продукти. Вони захищають інтереси творців, надаючи їм моральне та економічне право на свої твори. Згідно зі Всесвітньою організацією інтелектуальної власності (www.wipo.int), «під інтелектуальною власністю маються на увазі інформаційні матеріали або знання, що можуть бути також втілені у відчутних об’єктах, що, у свою чергу, можуть бути розповсюджені будь-де в світі у вигляді копій».

Існує дві галузі інтелектуальної власності: право на промислову власність та авторське право. Перше стосується патентів, що захищають винаходи, промислові зразки, торгові знаки та торгові назви. Авторське ж право стосується художньої творчості та технологічних проектів. Творець (автор) має певні права на свої твори (наприклад, запобігання його спотворення при копіюванні або адаптуванні). Інші права (наприклад, на копіювання) можуть використовуватися іншими персонами (наприклад, видавником), які отримали ліцензію від автора або носія авторського права. У багатьох випадках автори тестів (чи інших письмових робіт) передають авторське право видавнику або дистриб’ютору.

Оскільки навчальні та психологічні тести очевидно є продуктом людського розуму, вони захищені правами інтелектуальної власності. В багатьох випадках авторське право не покриває зміст питань чи завдань тесту (наприклад, ніхто не може мати права на таке завдання як «1+1 = (…)» чи твердження «Я відчуваю смуток»), але стосується самої структури тесту (структури шкал, системи балів, організації матеріалу і т.д.). Таким чином, імітування існуючого тесту, тобто збереження структури оригінального тесту з його системою балів, але створення нових завдань чи питань/тверджень – є порушенням права інтелектуальної власності. Якщо

тестовий розробник отримує авторське право для адаптування тесту, він (вона) має поважати оригінальні характеристики тесту (структуру, матеріал, формат, систему підрахунку балів), та змінювати їх лише у тому випадку, коли власник надає такий дозвіл.

Рекомендації для втілення. Розробники тестів мають поважати закон про авторське право та діючі погодження щодо використання оригінального тесту. Перед тим як розпочати адаптацію тесту, вони повинні отримати підписане погодження від носія права на інтелектуальну власність (тобто від автора або видавника). У договорі має бути зазначено, які саме модифікації характеристик оригінального тесту є прийнятними, а також чітко вказано, хто володітиме правами інтелектуальної власності на адаптовану версію.

PC-2 (2) Оцініть, чи в достатній мірі збігаються визначення та зміст вимірюваних тестом конструктів у цільових популяцій.

Пояснення. Даний принцип вимагає однакового розуміння вимірюваних конструктів серед різномовних та культурних груп, що є ключовим моментом валідних міжкультурних порівнянь. На цій стадії процесу тест або інший інструмент ще навіть не було адаптовано, так що бажано знайти та переглянути збірку емпіричних доказів, попередньо отриманих при дослідженні подібних тестів, а також судження з приводу збігання конструкту та завдань і їх придатності для залучених лінгвістичних та культурних груп. Зрештою дотримання цього важливого принципу має оцінюватися з допомогою емпіричних даних, що згадуються у переліку необхідних доказів в пункті С-2 (10). Метою будь-якого аналізу є не встановлення структури тесту (хоч це й побічний продукт будь-яких аналізів), але підтвердження еквівалентності структури тесту поміж його різномовних версій.

Рекомендації для практики. Для оцінки вимірюваного конструкту на його легітимність у контексті певної культурної/лінгвістичної групи треба найняти експертів, що спеціалізуються на даному конструкті і при цьому знайомі з культурними групами, які проходять даний тест. Для цього вони можуть спробувати відповісти на наступне питання: чи має сенс даний конструкт у контексті другої культури? Доволі часто нам доводилось спостерігати як комітет вирішував, що вимірюваний конструкт у, наприклад, навчальному тесті, має менше, або зовсім ніякого значення в іншій культурі (наприклад, якість життя, депресія чи інтелект). Завдяки фокус-групам, інтерв'ю та опитуванням можна отримати структуровану інформацію щодо ступеню збігу конструкту в обох культурах.

PC-3 (3). Мінімізуйте вплив будь-яких культурних або лінгвістичних відмінностей, що не мають відношення до передбачуваної мети використання тесту в цільових популяціях.

Пояснення. Культурні та лінгвістичні властивості, що не мають відношення до вимірюваних тестом перемінних, мають бути визначені на ранній стадії проекту. Вони можуть відноситись до формату завдань, матеріалів (наприклад, використання комп'ютеру, малюнків або ідеограм…), часових обмежень, і т.д.

Рекомендований підхід до вирішення цієї проблеми полягає в оцінці «лінгвістичної та культурної дистанції» між оригінальною та цільовою мовами та відповідними культурними групами. Оцінка лінгвістичної та культурної дистанції може включати в себе розгляд відмінностей в мові, структурі сім'ї, релігії, стилю життя та цінностях (van de Vijver & Leung, 1997).

Втілення даного принципу значним чином залежить від якісних методів, та експертів, знайомих із дослідженнями певних культурних та лінгвістичних відмінностей. Це значно ускладнює відбір перекладачів: вимагається, щоби вони були носіями мови перекладу та цільової культури, оскільки простого знання цільової мови недостатньо для виявлення можливих джерел методичної похибки. Наприклад, в китайсько-американському порівняльному дослідженні рівню математичних здібностей восьмикласників, проведеному Хамблтоном, Ю та Слейтером (Hambleton, Yu, and Slater, 1999), були виявлені проблеми формату та довжини тесту, разом із багатьма культурними особливостями, пов'язаними з математичним тестом для восьмикласників.

Рекомендації для втілення. Цей принцип є досить складним в плані реалізації за допомогою емпіричних даних. Його втілення особливо складне на ранніх етапах адаптації. Тим не менш, якісні дані завжди можуть бути зібрані таким чином:

- За допомогою спостережень, інтерв'ю, фокус-груп, чи опитувань, визначте мотиваційні рівні учасників, наскільки вони розуміють інструкції, попередній досвід складання психологічних тестів, швидкість виповнення завдань, ознайомленість із оціночними шкалами, та культурні відмінності (але навіть ці порівняння можуть виявитись проблематичними через культурні відмінності в плані розуміння саме змінних). Коли збір таких дослідницьких даних є проблематичним, отримайте всю можливу інформацію від перекладачів. Частково ця робота може бути виконана перед будь-яким прогресом в адаптації.
- Може бути можливим враховувати ці «прикрі змінні» в будь-яких послідуючих емпіричних аналізах як тільки тест буде адаптовано, і він стане придатним для дослідження валідності через аналіз коваріацій чи інші аналізи, що протиставляють різномовних та різнокультурних учасників між собою за такими змінними як мотиваційний рівень чи ознайомленість із певною шкалою балів (напр., Johnson, 2003; Javaras & Ripley, 2007).

Принципи розробки тесту

TD-1 (4) Обирайте спеціалістів із необхідним досвідом для гарантії того, що під час процесу перекладу та адаптації будуть прийняті до уваги лінгвістичні, психологічні та культурні відмінності обох популяцій.

Пояснення. Даний принцип залишив найбільш глибокий слід у сфері – існує вагоме свідчення того, що саме під його впливом тестові організації почали шукати перекладачів із кваліфікацією, не обмеженою лише тільки знанням двох мов, необхідних у процесі адаптування тесту (див., наприклад, Grisay, 2003). Знання культур та, як мінімум, базові знання у сфері тестування та створенні тестів стали невід'ємним критерієм для відбору перекладачів. Також відомо, що під впливом цього принципу організації, які займаються перекладом та адаптацією тестів, почали користуватися послугами якнайменш двох перекладачів – наприклад, для здійснення методу прямого та зворотнього перекладу. На сьогоднішній день минула практика доручати всю роботу одному перекладачеві вважається неприйнятною, навіть якщо рівень кваліфікації спеціаліста є дуже високим.

Знання та компетенція у сфері цільової культури гарантуються користуванням послугами перекладачів, які є носіями цільової мови як рідної та проживають у місцевості, населення якої даною мовою спілкується, при чому перша умова є необхідною, в той час як друга – більш ніж бажаною. Переклад носія цільової мови як рідної буде не тільки дуже точним, але й природнім та легким для сприйняття. Проживання в країні-носії цільової мови є гарантом найактуальніших знань про вживання сучасної мови.

Отже, в нашому розумінні «спеціаліст» є людиною із належними знаннями про (1) задіяні мови, (2) культури, (3) вміст тесту та (4) загальні принципи тестування, завдяки яким він або вона зможе виповнити переклад чи адаптацію тесту на високому, професійному рівні. На практиці може бути ефективніше залучати до роботи цілі команди, що складалися б із людей з різними кваліфікаціями (наприклад, з перекладачів зі спеціалізацією у певній сфері або без неї, тестового експерта і т.д.), для того, щоби виявляти ті моменти, які інші можуть прогледіти. В будь якому випадку, знання загальних принципів тестування разом із знанням тестового вмісту мають стати невід'ємною частиною навчальної підготовки перекладачів.

Рекомендації для втілення. Ми можемо порадити наступне:

- Обирайте перекладачів, що є носіями цільової мови як рідної та володіють поглибленими знаннями в сфері культури, до якої адаптується тест; також бажано, щоби вони проживали у країні-носії цільової мови та культури. Поширена помилка – визначати як перекладачів тих людей, що знають мову, але не дуже добре розуміються на культурі, оскільки поглиблені знання у сфері культури зазвичай є критичними для збереження культурної еквівалентності. Завдяки володінню такими знаннями можна з легкістю виявити культурні відсилки (наприклад, крикет, Ейфелева вежа, президент Лінкольн, кенгуру і т.д.), з якими місцеві працівники можуть бути незнайомі.
- По можливості обирайте перекладачів із досвідом у сфері, до якої відноситься вміст тесту, а також знанням принципів оцінювання (наприклад, у тестах множинного вибору вірна відповідь має бути не довшою чи коротшою за інші варіанти відповідей; граматичні формулювання не мають допомагати у пошуку правильної відповіді; а у відповідях типу «правда/неправда», твердження, що є правдивими, не мають бути помітно довшими за неправдиві).
- На практиці може бути майже неможливо знайти перекладачів, що знаються на принципах розробки тестів, тому необхідно провести спеціальний тренінг, щоби надати перекладачам уявлення про принципи написання завдань у тих форматах, з якими вони працюватимуть. У відсутність такого тренінгу занадто сумлінні з перекладачів можуть стати джерелом проблем, які можуть знизити валідність перекладеного тесту. Наприклад, іноді перекладач може додати роз'яснювальну примітку до відповіді, таким чином роблячи завдання простішим, ніж було задумано – у найбільш детально розписаному варіанті спокушені у складанні тестів учасники можуть впізнати правильну відповідь.

TD-2 (5) Використовуйте такі методи та проводьте такі процедури перекладу, щоби максимально збільшити придатність тестової адаптації для потрібних груп населення.

Пояснення. Даний принцип вимагає від перекладачів чи груп перекладачів прийняття таких рішень, що в результаті зробили би адаптовану версію максимально придатною для цільової

групи населення. Це означає, що мова має сприйматися природно, ясно та зрозуміло, і фокусуватися більшою частиною на функціональній, ніж буквальній еквівалентності. Для досягнення даних цілей найчастіше використовуються методи прямого та зворотнього перекладу. Тема даних методів достатньо повно розкрита у роботах Бріслін (Brislin, 1986) та Хамблтон і Пацули (Hambleton & Patsula, 1999): їх визначення, а також сильні та слабкі сторони. Слід зазначити, що обидва методи мають свої недоліки, а значить рідко коли можуть надати достатніх доказів для затвердження перекладеного та адаптованого тесту. Головним мінусом методу зворотнього перекладу є те, що у випадку, коли він здійснюється в найвузькішому сенсі, версія тесту цільовою мовою так і не стає об'єктом критичного огляду. В результаті можна отримати таку версію тесту цільовою мовою, яку достатньо легко перекласти у зворотньому напрямку, але сама по собі вона є досить незграбною.

Метод подвійного перекладу та процедура узгодження направлені на усунення недоліків та ризику разового перекладу. Цей підхід полягає в тому, що третій незалежний перекладач або група спеціалістів виявляє та вирішує будь-які розбіжності між альтернативними прямими перекладами, узгоджуючи їх і утворюючи єдину версію. У великомасштабних міжнародних оцінювальних програмах типу PISA дві різномовні версії (наприклад, англійська та французька) можуть бути використані в якості окремих джерел для перекладу, котрі пізніше узгоджуються в єдину версію цільовою мовою (Grisay, 2003). Цей підхід має серйозні переваги, включаючи виявлення можливих розбіжностей та виправлення їх безпосередньо в цільовій мові. До того ж, використання більш ніж однієї мови-джерела може допомогти зменшити вплив культурних властивостей оригіналу.

Відмінності у структурі мови можуть спричинити деякі проблеми під час перекладу тесту. Наприклад, у добре відомій шкалі, розробленій Роттером та Раффертi (Rotter and Rafferty 1950) англійською мовою, треба доповнити незавершені твердження в такому форматі: "I like….."; "I regret….."; "I can't…..". Однак цей формат є непридатним у турецькій мові, де додаток у реченні повинен стояти перед присудком та підметом. Використання незавершених речень як в оригіналі повністю змінило би образ надання відповідей, адже турецьким студентам довелося би спершу дивитися в кінець речення, щоби заповнити пропуск на початку.

Яким би не було вирішення цієї проблеми, версія тесту цільовою мовою так чи інакше відрізнялась би від оригінальної в плані особливостей формату.

Рекомендації для втілення. Для перевірки того, чи реалізовано даний принцип, не обійтись без збірки критичного матеріалу від рецензентів.

- Використовуйте оціночні шкали, розроблені Брісліном (Brislin, 1986), Джинрі та Бертрандом (Jeanrie & Bertrand, 1999), чи Хамблтоном та Зеніскі (Hambleton & Zenisky, 2010). Хамблтон і Зеніскі надають емпірично валідизований список із 25-ти різних характеристик перекладеного тесту, що мають бути перевірені під час процесу адаптування. Серед перевірочних питань від Хамблтона та Зеніскі (2010) є, наприклад, такі: *«Чи відповідає мова перекладеного завдання мові оригінального завдання тесту за своєю складністю та загальним характером?»* та *«Чи були внесені під час перекладу такі зміни (пропуски, заміни, чи додатки), що можуть вплинути на рівень складності завдання в обох версіях тесту?»*

- Використовуйте стільки методів перекладу, скільки можливо здійснити на практиці. Наприклад, метод зворотнього перекладу можна використати для повторної перевірки цільової версії, створеної за допомогою подвійного перекладу та узгодження, виповненого групою спеціалістів.
- Якщо передбачено використання тесту поміж різних культур, розгляньте варіант із синхронною/паралельною розробкою різномовних версій тесту з самого початку – для уникнення майбутніх проблем з перекладом/адаптацією оригінальної версії. Більше інформації щодо паралельної розробки тестів можна знайти, наприклад, у наступному джерелі -- Solano-Flores, Trumbull, and Nelson-Barber (2002). Як мінімум, створюйте таку оригінальну версію, що була б найбільш доступною для майбутніх перекладів, та дозволила б уникнути потенційних проблем наскільки це можливо – наприклад, обходячись без культурних відсилок, характерних форматів завдань та відповідей і т.д.
- Беручи до уваги синтаксичні відмінності серед різних мов, слід уникати використання таких форматів, що залежать від чіткої незмінної структури речення, у великомасштабних міжнародних оцінках та психологічних тестах – через можливі труднощі перекладу.

TD-3 (6) Надайте докази, що інструкції тесту та вміст завдань мають однакове значення для всіх цільових груп населення.

Пояснення. Докази, які треба мати згідно з даним принципом, можуть бути зібрані за допомогою здійснення певних стратегій (див., напр., van de Vijver and Tanzer, 1997). Серед таких стратегій: (1) залучення рецензентів-носіїв локальної культури та мови; (2) використання вибірок двомовних респондентів; (3) проведення локальних досліджень для оцінки тесту; та (4) використання нестандартних способів проведення тесту для подальшого підвищення ступеню його придатності та валідності.

Проведення маленького випробування адаптованої версії тесту – це хороша ідея. У таке випробування може входити не лише тільки проведення тесту та аналіз даних, але і – що найважливіше – інтерв'ю з інспекторами та протестованими для отримання зворотньої реакції. Також можливі такі способи, коли використовуються послуги контент-експертів, що є носіями різних мов та відповідних культур, або експертів-білінгвів. Наприклад, двомовних контент-експертів можна попрохати оцінити, наскільки подібні за своєю складністю формати завдань та вміст обох тестів. Багатообіцяючим є також метод когнітивного інтерв'ю (Levin, et al., 2009).

Рекомендації для втілення. Вище було надано декілька варіантів втілення даного принципу. Наприклад,

- Залучайте рецензентів-носіїв локальної культури та мови для оцінки перекладу/адаптації тесту.
- Використовуйте вибірки двомовних респондентів для оцінки еквівалентності обох версій тесту, їх інструкцій та завдань.
- Проводьте місцеві дослідження для оцінки тесту. Такі невеликі випробування можуть бути дуже цінними. Обов'язково опитайте інспектора та респондентів після проведення тесту, адже часто їх коментарі мають більше значення аніж відповіді саме на тестові завдання.

- Адаптуйте проведення тесту для підвищення придатності та валідності. Немає сенсу притримуватися таких самих тестових інструкцій, якщо респонденти не в змозі правильно їх зрозуміти на цільовій мові/у контексті локальної культури.

TD-4 (7) Надайте докази того, що формати завдань, оціночні шкали, категорії оцінок, тестові стандарти, способи проведення та інші процедури підходять для всіх цільових груп населення.

Пояснення. Такі формати завдань як п'ятибальна рейтингова шкала, чи інші нові формати типу «перетягування», або «оберіть всі правильні відповіді», або навіть «оберіть одну правильну відповідь» можуть бентежити тих респондентів, які ніколи раніше з такими форматами завдань не зустрічалися. Навіть схеми завдань, використання графічних засобів або найпрогресивніші комп'ютеризовані формати можуть викликати стан замішання в учасників тесту. У США, з їх прагненням перевести більшість стандартизованих тестів для дітей у комп'ютерний формат, можна знайти багато прикладів таких помилок. Це перестає бути проблемою для багатьох дітей після спеціальних практичних занять. Респонденти повинні бути знайомими з цими новими форматами, або ж слід представляти джерело похибок тестування, що можуть спотворити результати будь-яких індивідуальних або групових тестів.

Проблеми можуть виникати з версіями тесту, що проводяться на комп'ютері. Якщо респонденти не ознайомлені з платформою комп'ютеризованих тестів, слід проводити консультації, які б гарантували, що респонденти зможуть отримати інформативні результати.

Рекомендації для втілення. При оцінці того, чи було реалізовано даний принцип, слід опиратися як на кількісні, так і на якісні показники. Слід перевірити адаптований тест на наявність наступних рис:

- Упевніться, що практичні тренувальні заняття підходять для того, щоби респонденти змогли надати такі відповіді, які би в повній мірі відображали рівень їх знань/ здібностей/освоєння матеріалу.
- Упевніться, що респонденти знайомі з будь-якими новими форматами завдань або способами проведення тесту (наприклад, комп'ютеризованими), що складають процес тестування.
- Упевніться, що всі правила тестування (наприклад, розташування будь-яких об'єктів або позначення відповідей на бланку) зрозумілі респондентам.
- Скористуйтеся оціночними анкетами Джинрі та Бертранду (Jeanrie & Bertrand, 1999) або Хамблтона та Зеніскі (Hambleton & Zenisky, 2010). Наприклад, в анкеті Хамблтона та Зеніскі є такі питання: «Чи є однаковим формат завдань (включаючи їх розміщення) в обох версіях?» та «Якщо у завданні в оригінальній версії на якомусь слові чи фразі зроблено акцент (за допомогою жирного шрифту, підкреслення, курсиву і т.д.), то чи було зроблено те ж саме у перекладеній версії завдання?».

TD-5 (8) Зберіть експериментальні дані щодо адаптованого тесту для проведення аналізу завдань, оцінки надійності та невеликих досліджень валідності, щоби внести усі необхідні поправки до адаптованого тесту.

Пояснення. Перед тим як проводити великомасштабні дослідження надійності та валідності тестових результатів, що можуть бути досить коштовними та займати багато часу, слід отримати докази, підтверджуючі психометричну якість адаптованого тесту. Існує багато

способів психометричного аналізу, які дозволяють отримати докази надійності та валідності результатів вже на ранніх етапах. Наприклад, на етапі розробки тесту можна використати хоча б невелику вибірку (наприклад, зі ста осіб) для аналізу завдань, і це вже надасть таку важливу інформацію про функціонування окремих завдань тесту. Ті завдання, що виявляються надто легкими чи складними у порівнянні з іншими, або мають низький чи негативний показник дискримінантної сили, можна перевірити на предмет можливих вад. У випадку з завданнями із множинним вибором доцільно дослідити ефективність відповідей-дистракторів. Виявити проблеми та вирішити їх – цілком реально. До того ж, разом із даними, отриманими з аналізу завдань, коефіцієнти альфа та омега (McDonald, 1999) забезпечують розробника тесту усією важливою інформацією, що може бути використана для підтримки рішень щодо належної довжини тестових версій оригінальною та цільовою мовами.

В деяких випадках певні аспекти адаптування досі можуть викликати питання: чи будуть інструкції до тесту зрозумілими в повній мірі? Чи мають вони бути іншими, щоби ефективно направляти учасників тесту? Чи не будуть комп'ютеризовані тести представляти собою проблему для окремих респондентів (наприклад, з низьким соціально-економічним становищем) з цільової групи населення? Чи не занадто багато питань дається для виділеного часу? На всі ці та багато інших питань можна відповісти за допомогою невеликих досліджень валідності. Метою буде зібрати достатню кількість даних для прийняття рішення щодо того, рухатися далі із адаптованим тестом чи ні. Якщо було вирішено йти вперед, можна планувати та проводити ряд інших, істотно амбітніших досліджень (наприклад, дослідження рівню диференціального функціонування завдань [DIF] та факторної структури тесту).

Рекомендації для втілення. Існує ряд початкових аналізів, що можна провести:

- Проведіть традиційний аналіз завдань, щоби отримати інформацію про індекси дискримінативності завдань. Також проведіть аналіз дистракторів у випадку з завданнями множинного вибору і подібних.
- Проведіть аналіз надійності (наприклад, KR-20 у випадку з дихотомічною обробкою результатів, або коефіцієнт альфа чи коефіцієнт омега у випадку з політомною моделлю обробки).
- По мірі необхідності, проведіть 1-2 дослідження, щоби мати уявлення про валідність адаптованого тесту. Наприклад, припустимо, що адаптований тест має проводитися на комп'ютері. Бажано провести дослідження для оцінки способу проведення тесту (тобто, бланковий тест проти комп'ютеризованого). Припустимо, що інструкції вимагають відповісти на всі питання. Може бути важливо пошукати інформацію з приводу того, які інструкції підійдуть найкраще для цієї мети. Дослідники виявили, що заставити деяких респондентів відповісти на кожне питання може бути неочікувано складно, якщо заохочується вгадування у випадку нестачі необхідної інформації для відповіді у респондентів.

Принципи затвердження (емпіричного аналізу)

Принципи затвердження є тими, що ґрунтуються на емпіричному аналізі повномасштабних досліджень валідності.

C-1 (9) Обирайте вибірку з характеристиками, відповідними до передбачуваного використання тесту, а також належного розміру та релевантності для емпіричних аналізів.

Пояснення. Метод збору даних означає спосіб, в який збирається інформація для встановлення норм (якщо потрібно) та еквівалентності різномовних версій тесту, проведення досліджень валідності та надійності, а також диференціального функціонування завдань. Першою вимогою до збору даних є те, що вибірки мають бути достатньо великими задля можливості отримання стабільної статистичної інформації. Хоч ця вимога є ключовою для будь-якого типу дослідження, вона є особливо важливою в контексті дослідження валідності адаптації тесту, оскільки необхідні для встановлення еквівалентності тесту та завдань статистичні техніки (наприклад, конфірматорний факторний аналіз, методологія теорії відповідей на тестові завдання [IRT] для виявлення можливих похибок у завданнях тесту) можуть бути застосовані найконструктивнішим чином разом із вибірками, достатньо великими для надійної оцінки параметрів моделі (рекомендований розмір вибірки буде залежати від складності моделі та якості даних).

Також вибірка для повномасштабного дослідження валідності має бути репрезентативною ілюстрацією цільової групи населення, на яку зорієнтований тест. Пропонуємо вашій увазі важливу роботу ван де Війвера та Танцера (1997), а також методологічний вклад Хамблтона, Меренди та Шпільбергера (2005), Бірна (2008) та Бірна і ван де Війвера (2014) в якості керівництва з відбору належних статистичних планів та аналізів. Сіречі (Sireci, 1997) детально розглянув проблеми, виникаючі при поєднанні багатомовних тестів із звичайною шкалою.

Іноді виходить так, що цільова група населення, на яку зорієнтовано версію тесту новою мовою, отримує набагато менші чи більші результати, або є більш чи менш гомогенною ніж та група, якій призначався тест оригінальною мовою. Через це проблематично використовувати такі методи аналізу як дослідження надійності та валідності. Одним із варіантів рішення буде відібрати підвибірку з оригінальної популяції для порівняння з вибіркою цільової. За допомогою порівняльного дослідження можна усунути будь-які розбіжності у результатах вибірок, що могли виникнути внаслідок відмінностей у формах розподілу в обох групах (див., Sireci & Wells, 2010). Наприклад, порівняння структури тестів зазвичай включає в себе коваріацію, а вона буде варіюватися так само як і функція розподілу балів. Якщо використовувати спарені вибірки, яку б роль не грав розподіл оцінок в результатах, його буде зіставлено в двох вибірках, так що вплив розподілу на результати можна буде виключити з можливих пояснень відмінностей в результатах.

Можливо, ще один приклад допоможе розібратися в проблемі різниці розподілу балів в цільовій групі та групі-носії оригінальної мови. Припустимо, що коефіцієнт надійності результатів тесту дорівнює .80 в групі-носії оригінальної мови, в той час як коефіцієнт у групи-носія мови цільової – лише .60. Така різниця може викликати занепокоєння та сумніви у придатності версії тесту новою мовою. Тим не менш, часто до уваги не береться той факт, що надійність є сумісною властивістю тесту та популяції (McDonald, 1999), оскільки вона залежить як від дисперсії істинного балу (властивість популяції), так і від дисперсії похибки (властивість тесту). Тому та сама дисперсія похибки може привести до вищої надійності просто через більшу дисперсію істинного балу в групі-носії оригінальної мови. МакДональд (1999) показує, що стандартна помилка вимірювання (квадратний корінь дисперсії похибки) є величиною більш придатною для порівнянь між вибірками, ніж надійність. Іншим альтернативним

способом використання коефіцієнтів надійності буде виробити зрівняну вибірку протестованих з групи-носія цільової мови та перерахувати надійність тестових балів.

Сучасний підхід до дослідження інваріантності оцінювання з використанням конфірматорного факторного аналізу множинних груп дозволяє оцінювати вибірки з різним розподілом латентних властивостей. В таких моделях середнім показникам, дисперсіям та коваріаціям латентних рис дозволено варіюватися поміж груп, в той час як передбачається, що такі параметри вимірювання як навантаження факторів та пересічення є рівними в усіх групах. Це дозволяє застосовувати повні вибірки та надає найреалістичніший сценарій різних розподілень вимірюваних рис серед різних популяцій.

Рекомендації для втілення. Практично в усіх дослідженнях будуть актуальними дві вказівки для охарактеризування вибірки(ок).

- Зберіть вибірку настільки велику, наскільки це доцільно, беручи до уваги те, що для виявлення потенційних похибок в тестових завданнях необхідно як мінімум 200 чоловік на одну версію тесту (Mazor, Clauser & Hambleton). Для проведення аналізу сучасної теорії тестування та дослідження підбору моделі потребується вибірка з якнайменш 500 респондентів (Hulin, Lissak & Drasgow, 1982; Hamleton, Swaminathan & Rogers, 1991), тоді як для досліджень факторної структури тесту вимагаються достатньо великі розміри вибірки, можливо, 300 чи більше респондентів (Wolf, Harrington, Clark & Miller, 2013). Безсумнівно, також можливі й аналізи з меншими вибірками, але все одно по можливості слід виробляти великі вибірки.
- Обирайте по можливості репрезентативні вибірки респондентів. Узагальнення висновків з нерепрезентативних вибірок респондентів є обмеженими. Щоби позбутися відмінностей в результатах через методологічні фактори типу варіацій в розподілі балів, хорошою ідеєю буде вироблення вибірки з оригінальної популяції для спарювання з цільовою популяцією.

C-2 (10) Надайте належні статистичні докази еквівалентності конструктів, методів та завдань для усіх передбачених популяцій.

Пояснення. Встановлення еквівалентності конструктів версії оригінальною та цільовою мовами є важливим, але це не єдиний емпіричний аналіз, який необхідно провести. Також тема підходів до встановлення еквівалентності конструктів (PC-2) та методів (PC-3) вже підіймалась вище у цьому керівництві.

Окрім цього, дослідники мають приділяти увагу еквівалентності завдань. Дослідження еквівалентності завдань називають «аналізом диференціального функціонування завдань (DIF)». Загалом, DIF наявне якщо два учасника тесту, які відносяться до різних (в культурному та лінгвістичному плані) груп, мають однаковий рівень вимірюваної властивості, але різну вірогідність відповіді на тестове завдання. Загальні відмінності у виповненні тесту могли виникнути серед груп, але це само по собі не представляє проблему. Якщо члени популяції зіставлені на основі вимірюваного тестом конструкту (зазвичай зіставляються загальні тестові бали або загальний бал мінус оцінка за завдання-предмет дослідження), і видно, що присутні відмінності у виповненні завдання поміж груп, це значить, що DIF наявне у завданні. Такий аналіз проводиться з кожним завданням тесту. Пізніше робиться спроба зрозуміти можливі причини диференціального функціонування завдань, і на основі даного критичного розгляду

можуть бути зроблені висновки щодо дефектності деяких завдань, які пізніше заміняються або взагалі видаляються із тесту.

Необхідно оцінити два таких потенційних джерела диференціального функціонування, як складнощі перекладу та культурні відмінності. Якщо точніше, то DIF може виникнути через (1) нерівнозначність перекладу, виникаючу під час перекладу з оригінальної мови на цільову, через ступінь вживаності лексики, змін у складності завдань, змін в еквівалентності значень та ін., (2) культурні та контекстуальні відмінності (Scheuneman & Grima, 1997; van de Vijver & Tanzer, 1997; Ercikan, 1998, 2002; Allalouf, Hambleton, & Sireci, 1999; Sireci & Berberoğlu, 2000; Ercikan, et al., 2004; Li, Cohen, & Ibera, 2004; Park, Pearson & Reckase, 2005; and Ercikan, Simon, & Oliveri, 2013).

Під час перекладу існує можливість використання маловживаних у цільовій мові слів. Значення можуть збігатися в обох мовах, але в одній культурі слово може бути більш широковживаним та відомим у порівнянні з іншою культурою. Також, у результаті перекладу виникає можливість змінити рівень важкості завдання через зміну довжини речення та його складності, або ж через використання простих або складних слів. Сенс також може змінитися, якщо деякі частини речення були видалені, дещо було перекладено неточно, було вжито слів, що мають більше одного значення в цільовій мові, через нерівнозначне тлумачення значень слів в різних культурах тощо. Насамперед культурні відмінності впливають на диференціальне функціонування завдань у різномовних версіях. Наприклад, такі слова як «гамбургер» або «касовий апарат» можуть бути незрозумілі представникам іншої культури або ж мати інше значення в її контексті.

Існує як мінімум чотири групи аналізу, завдяки яким можна перевірити, чи не функціонують завдання різним чином з-поміж лінгвістичних та/або культурних груп. Серед них (a) методи на основі теорії відповідей на завдання (IRT) (див, наприклад, Ellis, 1989; Thissen, Steinberg, & Wainer, 1988; 1993; Ellis & Kimmel, 1992), (b) метод Мантеля-Хензеля (MH) (див. Dorans & Holland, 1993; Hambleton, Clauser, Mazor, & Jones, 1993; Holland & Wainer, 1993; Sireci & Allalouf, 2003), (c) методи логістичної регресії (LR) (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993), та (d) процедура обмеженого факторного аналізу (ОФА) (Oort & Berberoğlu, 1992).

В IRT-підходах учасники тесту (з обох культурних/лінгвістичних груп) об’єднуються на основі набраних з латентної риси балів. В MX- та ЛР-методах, тестові бали, що спостерігаються або оцінюються, використовуються як критерій зіставлення перед порівнянням виповнення завдань респондентами в двох групах. Хоч сумарний бал є найпопулярнішим критерієм відповідності в таких процедурах, можуть використовуватися й інші оцінювані бали, отримані з, наприклад, факторного аналізу. Критерій відповідності має бути достатньо валідним та надійним для належної оцінки DIF. В ОФА-методах завдання досліджуються на предмет групуючої змінної (потенційного порушника) та латентної риси. Кожне завдання звільняється від факторного навантаження, і придатність до моделі оцінюється на підставі недійсної моделі, де жодне завдання не завантажене за групуючою змінною (модель без DIF). Якщо модель демонструє значно кращий рівень узгодженості, це вказує на DIF завдання.

В тому випадку, коли тест є дімензіонально складним, знайти належний критерій для порівняння може бути проблематично (Clauser, Nungester, Mazor & Ripkey, 1996). Використання багатофакторного критерію відповідності як, наприклад, різні факторні бали, отримані в результаті факторного аналізу, може також змінити інтерпретацію рівня DIF. Таким

чином, якщо тест є багатофакторним, дослідники можуть використовувати різноманітні критерії, щоби помічати завдання як диференціально функціонуючі, та оцінювати завдання, помічені як такі, спираючись на різні критерії відповідності. Багатофакторна відповідність може зменшити кількість завдань, що демонструють DIF поміж лінгвістичних та культурних груп.

Ці методи можуть потребувати різних розмірів вибірки. Робота моделей МХ, ЛР та ОФА може бути цілком надійною та валідною навіть із відносно малими вибірками у порівнянні з IRT-техніками, що вимагають більших вибірок для валідних оцінок параметрів. Іншим аспектом для розгляду є тип даних про відповіді на завдання. МХ, ЛР та ОФА можна застосовувати з бінарними даними. Інші ж підходи, такі як узагальнений МХ, потрібні для політомних даних.

Даний принцип зобов'язує дослідників встановлювати місцезнаходження можливих джерел методичної похибки в адаптованому тесті. До джерел методичної похибки входить: (1) різні рівні мотивації проходження тесту в учасників, (2) диференціальний досвід респондентів із психологічними тестами, (3) вища швидкість складання тесту групою-носієм однієї мови у порівнянні з іншою, (4) диференціальна ознайомленість із форматом надання відповідей поміж різномовних груп, та (5) гетерогенність стилю надання відповідей тощо. Похибки у відповідях були, наприклад, джерелом значних проблем у тлумаченні результатів PISA, тому отримали багато уваги дослідників у свій час.

Нарешті, цей принцип зобов'яже дослідників звернути увагу на конструктну еквівалентність. Існує якнайменш чотири статистичних підходи до оцінки еквівалентності конструктів в оригінальній та цільовій версіях тесту: оцінка факторної структури (ОФС), конфірматорний факторний аналіз (КФА), багатофакторне шкалювання (БФШ) та порівняння номологічної мережі (Sireci, Patsula, & Hambleton, 2005).

Згідно з ван де Війвером та Пуртінга (1991), факторний аналіз (як ОФС, так і КФА) є найбільш популярним статистичним методом для оцінки еквівалентності конструктів в різних культурах. Це судження досі залишається вірним, хоч і статистичне моделювання значно розвинулось з тих пір (див., наприклад, Hambleton & Lee, 2013; Byrne, 2008). Оскільки за допомогою ОФС складно порівняти окремі факторні структури, і не існує жодних загальноприйнятих правил щодо того, які структури можуть вважатися еквівалентними, такі статистичні підходи як КФА (див., наприклад, Byrne, 2001, 2003, 2006, 2008) та зважене багатофакторне шкалювання (ЗБФШ) є більш бажаними, адже вони можуть одночасно узгоджувати декілька груп (Sireci, Harter, Yang, & Bhola, 2003).

Було проведено багато досліджень, в котрих КФА був використаний для оцінки узгодженості факторної структури оригінальної версії тесту серед усіх його адаптованих версій (наприклад, Byrne & van de Vijver, 2014). КФА є таким привабливим в якості інструменту оцінки структурної еквівалентності адаптованих тестів тому, що він може впоратися одночасно з множинними групами, при цьому доступні статистичні тести підбору моделі, а також надані описові індекси підбору моделі (Sireci, Patsula, & Hambleton, 2005). Здібність працювати з множинними групами є особливо важливою, оскільки адаптація тесту під багату кількість мов становиться звичайною практикою (наприклад, деякі інструменти з оцінки інтелекту тепер перекладено/адаптовано сотнею мов, а тести TIMSS та OECD/PISA адаптовано тридцятьма мовами). Так як в КФА є сувора вимога – ніякого розподілу навантаження, цей метод часто не підходить для даних складних багатофакторних інструментів, і тому зростає популярність

моделювання структурними рівняннями (МСР) як дослідницького, особливо при необхідності оцінки особистісних даних або більш складних взаємопов'язаних змінних (Asparouhov & Muthen, 2009).

МСР є іншим привабливим підходом до оцінки еквівалентності конструкту поміж різномовних версій тесту. Як і ОФС, аналіз за допомогою МСР не вимагає певної структури тесту і, як КФА, дозволяє аналізувати множинні групи (напр., Sireci, et al., 2003).

Ван де Війвер і Танцер (1997) запропонували, щоби міжкультурні дослідники досліджувати надійність кожної культурної версії тесту в пошуках як конвергентних, так і дискримінантних доказів валідності в кожній цільовій групі. Такі дослідження часто можуть бути практичнішими за дослідження тестової структури, що потребують досить істотних розмірів вибірок.

Однак слід визнати, що порівняння успішності виповнення різномовних версій тесту учасниками не завжди є метою перекладу/адаптації тесту. Мета може полягати, наприклад, лише в тому, щоби була можливість оцінювати учасників тесту різних лінгвістичних та культурних груп за одним конструктом. В такому випадку необхідне ретельне вивчення валідності тесту в групі-носії другої мови, але дослідження для пошуку доказів еквівалентності двох форм не таке критичне. Значимість цього принципу буде залежати від передбачуваної мети тесту цільовою мовою. Тести типу PISA та TIMSS потребують доказів значного співпадання вмісту, оскільки результати використовуються для порівняння навчальної успішності студентів з багатьох країн. Застосування шкали депресії, перекладеної з англійської на китайську, з метою дослідження депресії або, наприклад, вимірювання її у клієнтів психологів, не потребує великого збігу контенту. Замість цього необхідність буде у валідності шкали депресії в Китаї.

Даний принцип також можна втілювати за допомогою статистичних методів вже після того, як тест було адаптовано. Наприклад, якщо підозрюється, що різнокультурні групи відрізняються за важливими змінними, що не відносяться до вимірюваного конструкту, такі «незручні» змінні можуть контролюватися за допомогою комплексних підходів та статистичних аналізів. Аналіз коваріації, рандомізований блочний план, та інші статистичні методи (регресійний аналіз, часткова кореляція тощо) можуть бути використані для регулювання впливу небажаних джерел варіації поміж груп.

Рекомендації для втілення. Цей принцип є дуже важливим, так що можна провести багато аналізів та досліджень. Для аналізів еквівалентності ми можемо запропонувати наступні вказівки:

- Якщо розміри вибірок такі, як треба, проведіть порівняльні дослідження еквівалентності конструкту тестових версій оригінальної та цільовою мовами. Існує немало пакетів програмного забезпечення для полегшення таких аналізів (див. Byrne, 2006).
- Проводьте оцінку факторної структури чи конфірматорний факторний аналіз, та/або аналіз зваженого багатофакторного шкалювання для встановлення рівня узгодженості структури певного тесту поміж різномовних та різнокультурних груп. Необхідною умовою є великі розміри вибірок (10 осіб на змінну), що ускладнює проведення цих дослідів в багатьох міжкультурних дослідженнях. Відмінну модель для дослідження такого типу представили Бірн та ван де Війвер (2014).

- Шукайте докази конвергентної та дискримінантної валідності (в першу чергу знайдіть кореляційні дані серед набору конструктів та перевірте стабільність цих кореляцій серед лінгвістичних та/або культурних груп) (див. van de Vijver & Tanzer, 1997).

Для аналізів диференціального функціонування завдань є декілька вказівок, вказаних нижче. Складніші підходи слід шукати в професійній літературі на тему DIF:

- Проводьте аналізи DIF, використовуючи одну зі стандартних процедур (при бінарній моделі обробки завдань найбільш підходящим буде метод Мантелю-Хензелю; якщо ж результати обробляються за допомогою політомної моделі, узагальнений метод Мантелю-Хензелю представлений як можливий варіант). Інші рішення, більш громіздкі та тяжчі, включають підходи, що базуються на IRT. Якщо розміри вибірок скромніші, дельта-функція може виявити потенційно дефектні завдання. Ще одним можливим варіантом є умовні порівняння (приклади порівняння результатів за допомогою методів, що не вимагають великих вибірок, можна знайти, наприклад, у Muñiz, Hambleton, & Xing, 2001).

C-3 (11) Надайте докази норм, надійності та валідності адаптованої версії тесту в передбачуваних популяціях.

Пояснення. Норми та докази валідності і надійності оригінальної версії тесту не можуть автоматично застосовуватися до інших можливих версій, перекладених на інші мови та адаптованих під інші культури. Саме тому необхідні емпіричні докази валідності та надійності будь-яких нових версій. Всі емпіричні докази, підтверджуючі висновки, зроблені на підставі тесту, мають міститися у довідникові користувача тесту. Особливу увагу слід приділити п’яти джерелам доказів валідності, що засновуються на: тестовому змісті, процесу надання відповідей, внутрішній структурі, взаємозв’язках між змінними, та наслідках тестування (AERA, APA, NCME, 2014). Оцінка факторної структури та конфірматорний факторний аналіз, моделювання структурними рівняннями та аналізи типу «множинні ознаки – множинні методи» -- це деякі зі статистичних методів, заснованих на внутрішній структурі, що можуть використовуватися в цілях отримання та аналізування даних для вирішення питання доказу валідності.

Рекомендації для втілення. Рекомендації такі ж що і до інших тестів, використання яких передбачається:

- Якщо передбачається, що норми, розроблені для оригінальної версії тесту, будуть використовуватися з адаптованою версією, слід надати докази того, що це цілком доречно та чесно з точки зору статистики. Вразі, коли не може бути представлено жодних доказів, які б виправдали використання оригінальних норм, треба розробити норми, специфічні для адаптованої версії, спираючись на стандарти розробки норм.
- Зберіть достатню кількість доказів надійності для виправдання використання версії тесту цільовою мовою. Як правило, в ряд доказів має входити оцінка внутрішньої узгодженості (напр., KR-20, або коефіцієнти альфа чи омега).
- Зберіть необхідну кількість доказів валідності для вирішення, чи має використовуватися версія тесту цільовою мовою. Тип зібраних доказів буде залежати від передбачуваного використання результатів (наприклад, змістовна валідність для тестів досягнень, прогностична валідність для тестів на професійну придатність і т.д.).

C-4 (12) Використовуйте належні техніки прирівнювання та методики аналізу даних коли зв'язуєте шкали оцінок з різномовних версій тесту.

Пояснення. При зв'язуванні двох різномовних версій тесту з єдиною звітною шкалою доступні декілька варіантів. Якщо використовується звичайний ряд завдань, їх функціонування має бути оцінено в обох лінгвістичних групах, і вразі виявлення диференціального функціонування треба розглянути видалення їх з даних, що використовуються для встановлення зв'язку. Дельта-графік (Angoff & Modu, 1973) добре справляється з цією задачею, і Кук та Шміт-Каскаллар (Cook & Shmitt-Cascallar, 2005) вдало проілюстрували в своїй праці як використати дельта-графік для виявлення завдань, що мають різне значення для двох груп екзаменованих. Не всі типи завдань мають однаковий потенціал для зв'язування різномовних версій. Оцінки складності завдання та показника відмінностей, виведені на основі загальних принципів IRT для звичайних завдань, можуть бути нанесені на графік, щоби допомогти виявити звичайні працюючі неналежним чином завдання (див. Hambleton, Swaminathan, & Rogers, 1991).

Але зв'язування (тобто «прирівнювання») оцінок з двох різномовних версій тесту завжди буде проблематичним, оскільки треба зробити сильні припущення про дані. Іноді робиться дуже проблематичне припущення про те, що різномовні версії тесту є еквівалентними, а значить оцінки з двох версій тесту є взаємозамінними. Таке припущення могло б заслуговувати увагу у випадку з математичними тестами, тому що переклад/адаптація, як правило, є досить простими та дохідливими. Воно також могло б бути перспективним, якщо дві версії тесту були ретельно сконструйовані, так що можна було зробити припущення, що оригінальна версія тесту працює з відповідною популяцією рівнозначно тому, як цільова версія працює з цільовою популяцією. Воно було б перспективним, якщо всі інші доступні дані вказували на те, що дві різномовні версії тесту є еквівалентними, і немає жодних методичних похибок, що впливають на результати версії тесту цільовою мовою.

Існує ще два рішення, жодне з яких ідеальне. По-перше, зв'язування могло б бути зроблено з підвибіркою завдань, що вважаються «суттєво еквівалентними» в обох версіях тесту. Наприклад, завдання можуть бути такими, які оцінили як дуже прості для перекладу/адаптації. В принципі, рішення могло б спрацювати, але воно вимагає зв'язування завдань та вимірювання залишком тестових завдань єдиного конструкту. Друге рішення включає в себе зв'язування за допомогою вибірки учасників тесту-білінгвів. Завдяки складанню обох версій тесту цією вибіркою стало б можливо створити таблицю конвертації оцінок. Вибірка не могла б бути дуже малою, а порядок презентації форм тесту був би врівноваженим. Великим припущенням в даному підході є те, що респонденти – дійсно білінгви, а значить, не дивлячись на відносні складнощі в формах, вони мають скласти обидві форми однаково гарно. Будь-яка відмінність буде використана для коригування балів в процесі конвертації балів з однією версії тесту в іншу.

Рекомендації для втілення. Зв'язування оцінок між адаптованими версіями тесту буде в кращому випадку проблематичним, тому що всі техніки прирівнювання мають якнайменш один великий недолік. Скоріш за все, найкращою стратегією було б направити зусилля на дотримання всіх заходів встановлення еквівалентності оцінок. Якщо дані відповідають усім трьом, наведеним нижче, питанням, оцінки двох версій тесту навіть можуть вважатися рівнозначними:

- Чи наявні докази того, що той самий конструкт вимірюється в оригінальній та цільовій версіях тесту? Чи так само пов'язаний даний конструкт з іншими зовнішніми змінними в новій культурі?
- Чи наявні вагомі докази того, що джерела методичної похибки були вилучені з тесту (наприклад, жодних проблем, пов'язаних із часом, використовувані формати знайомі всім учасникам, інструкції цілком зрозумілі, жодних випадків систематично невірних даних в одній чи іншій групі, стандартизовані інструкції, відсутність стилів відповіді (диференціальна мотивація, надто високі чи низькі оцінки в завданнях з рейтингами…)?
- Чи не містить тест завдань із можливою похибкою? Тут може стати в нагоді p-значення чи, ще краще, величина коефіцієнту дельта із завдань обох версій тесту. Бали, що не знижуються наряду з кривою рівняння першого ступеню мають бути досліджені на предмет придатності завдань, до яких вони відносяться, до використання в обох мовах. Аналізи DIF надають навіть вагоміші докази еквівалентності завдань поміж різномовних та різнокультурних груп.
- Якщо планується зв'язування балів, треба обрати та здійснити належний для цього метод. Також повинні бути надані дані, підтверджуючі валідність методу.

Принципи проведення тестування

A-1 (13) Підготовлюйте матеріали та інструкції з проведення тесту для мінімізації будь-яких проблем, пов'язаних із культурою та мовою, що можуть бути спричинені такими процедурами проведення тесту, які впливають на валідність зроблених на підставі балів висновків.

Пояснення. Реалізація керівних принципів проведення тесту має починатися з аналізу всіх факторів, що можуть загрожувати валідності тестових балів в певному культурному та лінгвістичному контексті. Наявність досвіду проведення тесту в контексті єдиної культури та одномовності вже може допомогти передбачити можливі проблеми. Наприклад, досвідчені інспектори тестування як правило знають, які аспекти інструкції можуть бути складними для респондентів. Ці аспекти можуть залишитися складними і після перекладу або адаптування. Також може бути так, що при застосуванні інструментів в новому лінгвістичному або культурному контексті виникнуть проблеми, які не спостерігалися раніше при проведенні тесту в контексті єдиної культури.

Рекомендації для втілення. Необхідно передбачити потенційні фактори, що можуть створити проблеми під час проведення тесту. Ось деякі з факторів, які необхідно дослідити задля забезпечення більшого ступеню справедливості та ефективності проведення тесту:

- Чіткість та зрозумілість тестових інструкцій (включаючи переклад цих інструкцій), спосіб надання відповідей (наприклад, заповнення бланку для відповідей), допустимі терміни складання тесту (однією з основних помилок є виділення часу, недостатнього для завершення всіх завдань), зацікавленість екзаменованих у проходженні тесту, обізнаність щодо мети тестування та способу підрахунку тестових балів.

A-2 (14) Точно визначайте, яких умов треба строго дотримуватися в усіх досліджуваних популяціях.

Пояснення. Мета даного принципу полягає в тому, щоби заохотити розробників тестів видавати інструкції з тестування та споріднених процедур (напр., умови тестування, часові обмеження тощо), яким можна було б чітко слідувати в усіх досліджуваних популяціях. Першочергово призначення цього принципу полягає в тому, щоби сприяти дотриманню стандартизованих інструкцій інспекторами тестування. У той ж час може бути чітко встановлено умови, що потрібно створити для учасників з обмеженими можливостями. Наприклад, додатковий час, крупніший шрифт, особливо тихі умови проведення тесту і т.д. На сьогоднішній день у сфері тестування це відомо як «пристосування тесту». Метою пристосування є не підвищення балів учасників тесту, а створення таких умов тестування для даної групи екзаменованих, щоби вони могли повноцінно виконувати завдання тесту, демонструючи, що вони знають та вміють робити.

Мають бути зазначені варіації стандартизованих умов тестування, щоби пізніше, у процесі, можна було розглянути їх та їх вплив на генералізацію та тлумачення.

Рекомендації для втілення. Даний принцип може частково перетинатися із A-1 (13), але тут він повторюється для підкреслення важливості складання учасниками тесту в настільки рівних умовах, наскільки цього можливо досягнути. Це необхідно, якщо передбачається взаємозамінне використання результатів двох різномовних версій. Ось деякі вказівки:

- Інструкції з тестування та споріднених процедур мають бути адаптовані та переписані в стандартизований спосіб, придатний для нової мови та культури.
- Якщо інструкції з тестування та споріднених процедур змінені та адаптовані під нові культури, інспектори мають навчитися новим процедурам; їх слід інформувати про ці процедури, а не оригінальні.

Принципи підрахунку балів та тлумачення

SS-1 (15) Тлумачте будь-які відмінності в групових результатах спираючись на всю доступну актуальну інформацію.

Пояснення. Навіть якщо тест було адаптовано за допомогою технічно обґрунтованих методів, та валідність тестових балів вже в якійсь мірі встановлена, слід мати на увазі, що значення внутрішньогрупових відмінностей може тлумачитися яким завгодно чином через культурні або інші відмінності між країнами або культурами, що приймали участь в тестуванні. Сіречі (2005) вніс поправки в підхід до оцінки еквівалентності двох різномовних версій тесту. Зміна полягала в проведенні обох мовних версій тесту окремо в групі учасників-білінгвів (тих, хто володіє обома мовами на високому рівні), які, до того ж, є носіями однієї культури та мови. Він виділив декілька варіантів дослідження із залученням респондентів-білінгвів, перелічив можливі сторонні змінні, які треба контролювати, та запропонував деякі цінні поради з приводу тлумачення здобутих відомостей.

Рекомендації для втілення. Є одна порада щодо того, як краще втілювати цю практику.

- В залежності від предмету дослідження (чи контексту, в рамках котрого проводиться порівняння груп), може бути розглянуто ряд можливих інтерпретацій перед встановленням однієї. Наприклад, слід виключити фактор диференціальної мотивації перед тим як робити висновок, що одна група впоралась краще за іншу. Контекст також

може значно впливати на виповнення тесту. Наприклад, індивіди однієї групи можуть продемонструвати гірші результати через навчання в умовах менш ефективної системи освіти.

SSI-2 (16) Зіставляйте бали популяцій лише коли встановлено рівень інваріантності в звітній шкалі.

Пояснення. Якщо основна увага приділяється порівняльним дослідженням серед лінгвістичних та культурних груп, оцінки, отримані при проведенні багатомовного тесту, мають бути зведені до загальної шкали. Цей процес називають «прирівнюванням» або «зв'язуванням». Для цього потребуються істотні розміри вибірок, а також докази того, що адаптована версія тесту не містить інструментальних та методичних похибок.

Ван де Війвер та Пуртінга (2005) описали декілька рівнів еквівалентності тесту поміж групносіїв різних мов та культур – їх робота краще за все допоможе розібратися в цьому понятті; більш того, саме вони і є його родоначальниками. Наприклад, вони відмітили, що еквівалентність одиниць вимірювання вимагає однакової метрики в звітних шкалах кожної групи, оскільки це гарантує те, що відмінності між людьми всередині груп мають одне й те саме значення (наприклад, відмінності між чоловіками та жінками з вибірки китайців можуть бути порівняні з вибіркою французів). Однак достовірне пряме зіставлення оцінок може проводитися, коли вони демонструють найвищий рівень еквівалентності, що називають скалярною еквівалентністю, яка вимагає наявності однакової одиниці вимірювання та однакового походження в групах.

Пропонується багата кількість методів (в рамках як класичної теорії тестування, так і теорії відповідей на завдання) для прирівнювання або зв'язування оцінок, отриманих з двох груп (або різномовних версій тесту). Зацікавлені читачі можуть звернутися до робіт Ангоффа (Angoff, 1984) та Колена і Бреннана (Kolen and Brennan, 2004) для більшого занурення в предмет. Кук та Шмітт-Каскалар (Cook and Schmitt-Cascallar, 2005) пропонують необхідний фундамент знань для розуміння доступних на даний момент статистичних методів для прирівнювання та шкалювання навчальних і психологічних тестів. Автори критикують певні методи зв'язування шкал, що використовуються в дослідженнях адаптованих тестів, та наводять приклади відібраних методів зв'язування і пов'язаних з ними проблем, описуючи три дослідження, що проводились за останні двадцять років для того, щоби зв'язати оцінки тесту *Scholastic Assessment Test* з *Prueba te Aptitude Academica*.

Рекомендації для втілення. Ключовим моментом тут є уникання надмірної інтерпретації тестових оцінок:

- Тлумачте результати, спираючись на доступні докази валідності. Наприклад, не треба проводити порівняльний аналіз рівнів успішності проходження тесту респондентів з двох лінгвістичних груп в тому випадку, якщо не була встановлена інваріантність виміру.

Принципи документування

Doc-1 (17) Надавайте технічну документацію з будь-якими змінами, враховуючи докази еквівалентності, коли тест адаптується під нову популяцію.

Пояснення. Великою кількістю дослідників (див., наприклад, Grisay, 2003) було усвідомлено та не раз підкреслено, наскільки важливим є даний принцип. TIMSS та PISA досягли значного успіху в дотриманні цього принципу, ретельно документуючи всі зроблені під час адаптації зміни. Маючи цю інформацію, можна зосередитись на доцільності зроблених змін.

Технічна документація також має містити всі необхідні деталі методології для майбутніх дослідників, щоби ті мали можливість повторити певні процедури, працюючи з тією ж чи іншою лінгвістичною та культурною групою. В ній має бути достатньо інформації з приводу доказів еквівалентності конструктів та шкалювання (якщо таке проводилося) як підтвердження придатності для використання в новій культурній групі. Якщо треба провести внутрішньопопуляційні порівняння, документація має містити звіт про дані, які використовувалися для порівнювання оцінок обидвох популяцій.

Іноді виникають питання щодо цільової аудиторії технічної документації. Документи мають вестись для технічного спеціаліста та людей, які будуть оцінювати придатність та практичну користь тесту для нових чи інших популяцій (для неспеціалістів можна робити короткий додатковий документ).

Рекомендації для втілення. До адаптованих тестів мають бути вироблені технічні керівництва з інформацією про всі якісні та кількісні дані, що відносяться до процесу адаптування. Особливо корисно фіксувати всі зміни, що були зроблені для акомодації тесту до цільової культури та мови. Здебільшого технічні спеціалісти та редактори журналів можуть захотіти продивитись документацію з описом створення та валідизації версії тесту цільовою мовою. Також вони, звичайно, побажають ознайомитись із результатами всіх аналізів. Слід торкнутися наступних питань:

- Чи наявні докази, підтверджуючі практичну користь і придатність конструкту та адаптованого тесту до вимірювань в новій популяції?
- Які дані про тестові завдання були зібрані, та з яких вибірок?
- Які ще дані були отримані для оцінки змістової, критеріальної та конструктної валідності?
- Яким чином аналізували різноманітні дані?
- Якими були результати?

Doc-2 (18) Забезпечте користувачів тесту документами, що сприятимуть успішному застосуванню адаптованого інструмента у контексті нової культури.

Пояснення. Цільовою аудиторією документації мають бути люди, що використовуватимуть тест для практичного оцінювання. Документи повинні відповідати принципам передової практики, викладеним у «Керівництві МТК з використання тестів» (див. www.InTestCom.org).

Рекомендації для втілення. Розробник тесту повинен надати чітку інформацію з приводу того, яким чином соціокультурні та екологічні фактори можуть вплинути на виконання тесту цільовою популяцією. Довідник користувача має:

- Описувати вимірювані у тесті конструкти та резюмувати дану інформацію; надавати опис процесу адаптації.

- Підсумовувати докази, що підтверджують факт адаптації, включаючи докази відповідності змісту завдань, інструкцій тесту, формату відповідей тощо культурним особливостям.
- Містити інформацію про придатність тесту до використання у роботі з різноманітними підгрупами всередині популяції, а також про обмеження його застосування.
- Охоплювати будь-які проблеми, що треба розглянути для успішного проведення тесту.
- Пояснювати, чи можуть проводитися внутрішньопопуляційні порівняння, і якщо можуть, то як саме.
- Забезпечувати всією необхідною інформацією про підрахунок балів та нормування (наприклад, про відповідні таблиці пошуку) або поясняти, як користувачі можуть отримати доступ до процедур підрахунку (якщо вони комп'ютеризовані).
- Надавати вказівки стосовно тлумачення результатів, включно з інформацією щодо того, як дані про валідність та надійність можуть впливати на зроблені на підставі тестових оцінок висновки.

ПІСЛЯМОВА

З нашого боку було докладено всіх зусиль, щоби створити такі керівні принципи, які змогли б допомогти розробникам та користувачам тестів в їх професійній практиці. Однак щоби змінити практику перекладу та адаптації на краще, одних лише принципів недостатньо – необхідно створити ефективні механізми їх розповсюдження. Нещодавній систематичний огляд Ріоса та Сіречі (Rios and Sireci, 2014) показав, що більшість опублікованих проектів з адаптації тестів не відповідали Керівництву МТК, котре існує вже близько 20 років. Саме тому ми заохочуємо читачів робити все можливе для того, щоби підвищити рівень обізнаності серед колег, інформуючи їх про другу редакцію даного Керівництва як про першоджерело передової практики, розробленої за допомогою професіоналів з усіх кутків світу.

Водночас ми усвідомлюємо, що друга редакція буде замінена наступною так само як і перша. Загальновідомі AERA, APA та NCME випустили вже шосту редакцію тестових стандартів (AERA, APA, & NCME, 2014). Очікується, що дане Керівництво МТК з адаптації тестів також зазнає перегляду в майбутньому. Якщо Вам відомі будь-які наукові статті чи роботи, які було б варто процитувати, або Ви бажаєте додати нові принципи чи внести поправки до існуючих вісімнадцяти – просимо повідомити про це МТК. Ви можете зв’язатися з діючим головою Комітету з досліджень та складання керівництв, що випустив другу редакцію, або ж із секретарем МТК по e-mail адресі, яку можна знайти за цим посиланням: www.InTestCom.org.

БІБЛІОГРАФІЯ

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Angoff, W. H., & Modu, C. C. (1973). Equating the scales of the Prueba de Apititud Academica and the Scholastic Aptitude Test (Research Rep No. 3). New York: College Entrance Examination Board.

Asparouhov, T. & Muthén, B. (2009). Exploratory structural modeling. *Structural Equation Modeling, 16*, 397-438.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137-164). Newbury Park, CA: Sage Publications.

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing, 1*, 55-86.

Byrne, B. (2003). Measuring self-concept measurement across culture: Issues, caveats, and application. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), *International advances in self research*. Greenwich, CT: Information Age Publishing.

Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Publishers.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*, 872-882.

Byrne, B. M., & van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10,* 107-132.

Byrne, B. M., & van de Vijver, F.J.R. (2014). Factorial structure of the Family Values Scale from a multilevel-multicultural perspective. *International Journal of Testing, 14*, 168-192.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripley, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214. Translating and Adapting Tests (Second edition) | **Final Version** | v.2.4

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139-170).

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and Practice* (pp. 137-166).

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology, 74*, 912-921.

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology, 77*, 177-184.

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research, 29*(6), 543-533.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3), 199-215.

Ercikan, K., Gierl, J. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321.

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *An international handbook for large-scale assessments* (pp. 110-124). New York:

Grégoire, J., & Hambleton, R. K. (Eds.). (2009). Advances in test adaptation research [Special Issue]. *International Journal of Testing, 9*(2), 73-166.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225-240.

Hambleton, R. K. (2002). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127-240.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1-16.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18. Translating and Adapting Tests (Second edition) | **Final Version** | v.2.4

Hambleton, R. K., & Lee, M. (2013). Methods of translating and adapting tests to increase cross-language validity. In D. Saklofske, C. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child assessment* (pp. 172-181). New York: Oxford University Press.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum Publishers.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications.

Hambleton, R. K., Yu, L., & Slater, S. C. (1999). Field-test of ITC guidelines for adapting psychological tests. *European Journal of Psychological Assessment, 15* (3), 270-276.

Hambleton, R. K., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 46-74). New York, NY; Cambridge University Press.

Harkness, J. (Ed.). (1998). *Cross-cultural survey equivalence*.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Javaras, K. N., & Ripley, B. D. (2007). An 'unfolding' latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*, 454-463.

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission Guidelines: Keeping validity in mind. *European Journal of Psychological Assessment, 15*(3), 277-283.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika, 68*, 563-583.

Kolen, M. J., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Stapleton Kudela, M., Stark, D., & Thompson, F. E. (2009). Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods, 3*(1), 13-25.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115-135.

Mazor, K.H., Clauser, B.E., & Hambleton, R.K. (1992). The effect of simple size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451. Translating and Adapting Tests (Second edition) | **Final Version** | v.2.4

Muniz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traduccion y adaptacion de los tests: segunda edicion. *Psicothema, 25(2),* 149-155.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115-135.

Oort, F. J., & Berberoğlu, G. (1992). Using restricted factor analysis with binary data for item bias detection and item analysis. In T. J. Plomp, J. M. Pieters, & A. Feteris (Eds.), *European Conference on Educational Research: Book of Summaries* (pp. 708-710). Twente, the Netherlands: University of Twente, Department of Education.

Park, H., Pearson, P. D., & Reckase, M. D. (2005). Assessing the effect of cohort, gender, and race on DIF in an adaptive test designed for multi-age groups. *Reading Psychology, 26*, 81-101.

Rios, J., & Sireci, S. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing, 14*(4), 289-312.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel- Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17*(2), 105-116.

Rotter, J.B. & Rafferty, J.E. (1950). *Manual: The Rotter Incomplete Sentences Blank: College Form*. New York: Psychological Corporation.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and Black examinees. *Applied Measurement in Education, 10*(4), 299-319.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16*, 12-19.

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Mahwah, NJ: Lawrence Erlbaum Publishers.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148-166.

Sireci, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*(3), 229-248.

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. Merenda, & C. Spielberger, C. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-116). Mahwah, NJ: Lawrence Erlbaum Publishers.

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*(2), 129-150. Translating and Adapting Tests (Second edition) | **Final Version** | v.2.4

Sireci, S. G., & Wells. C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33-68). Washington, DC: Council of Chief State School Officers.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-129.

Subok, L. (2017). Detecting differential item functioning using the logistic regression procedure in small samples. *Applied Psychological Measurement, 41(1)*, 30-43.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for Test Adaptation. *European Journal of Psychological Assessment, 15*, 258-269.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147- 169). Mahwah, NJ: Lawrence Erlbaum Publishers.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67-113). Mahwah, NJ: Lawrence Erlbaum Publishers.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89-99.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage Publications.

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*, 33-51.

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-308). Dordrecht, the Netherlands: Kluwer Academic Publishers.

van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17-24.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*, 29-37.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodical issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and* Translating and Adapting Tests (Second edition) | **Final Version** | v.2.4 *psychological tests for cross-cultural assessment* (pp. 39-64). Mahwah, NJ: Lawrence Erlbaum Publishers.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47*(4), 263-279.

Wolf, E.J., Harrington, K.M., Clark, S.L., & Miller, M.W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73(6),* 913–934.

ДОДАТОК А. Чек-лист Керівництва МТК з перекладу та адаптації тестів

Це чек-лист для того, щоби нагадати Вам про всі вісімнадцять принципів Керівництва МТК. Ми рекомендуємо помічати ті, що Вам уже вдалося успішно втілити в своєму проекті з перекладу/адаптації, і фокусуватися на тих, що ще не було реалізовано.

Принципи передумови

[ ] PC-1 (1) Перед здійсненням адаптації отримайте необхідний дозвіл від носія прав інтелектуальної власності на тест.

[ ] PC-2 (2) Оцініть, чи в достатній мірі збігаються визначення та зміст вимірюваних тестом конструктів у цільових популяцій.

[ ] PC-3 (3). Мінімізуйте вплив будь-яких культурних або лінгвістичних відмінностей, що не мають відношення до передбачуваної мети використання тесту в цільових популяціях.

Принципи розробки тесту

[ ] TD-1 (4) Обирайте спеціалістів із необхідним досвідом для гарантії того, що під час процесу перекладу та адаптації будуть прийняті до уваги лінгвістичні, психологічні та культурні відмінності обох популяцій.

[ ] TD-2 (5) Використовуйте такі методи та проводьте такі процедури перекладу, щоби максимально збільшити придатність тестової адаптації для потрібних груп населення.

[ ] TD-3 (6) Надайте докази, що інструкції тесту та вміст завдань мають однакове значення для всіх цільових груп населення.

[ ] TD-4 (7) Надайте докази того, що формати завдань, оціночні шкали, категорії оцінок, тестові стандарти, способи проведення та інші процедури підходять для всіх цільових груп населення.

[ ] TD-5 (8) Зберіть експериментальні дані щодо адаптованого тесту для проведення аналізу завдань, оцінки надійності та невеликих досліджень валідності, щоби внести усі необхідні поправки до адаптованого тесту.

Принципи затвердження

[ ] C-1 (9) Обирайте вибірку з характеристиками, відповідними до передбачуваного використання тесту, а також належного розміру та релевантності для емпіричних аналізів.

[ ] C-2 (10) Надайте належні статистичні докази еквівалентності конструктів, методів та завдань для усіх передбачених популяцій.

[ ] C-3 (11) Надайте докази норм, надійності та валідності адаптованої версії тесту в передбачуваних популяціях.

[ ] C-4 (12) Використовуйте належні техніки прирівнювання та методики аналізу даних коли зв'язуєте шкали оцінок з різномовних версій тесту.

Принципи проведення тестування

[ ] A-1 (13) Підготовлюйте матеріали та інструкції з проведення тесту для мінімізації будь-яких проблем, пов'язаних із культурою та мовою, що можуть бути спричинені такими процедурами проведення тесту, які впливають на валідність зроблених на підставі балів висновків.

[ ] A-2 (14) Точно визначайте, яких умов треба строго дотримуватися в усіх досліджуваних популяціях.

Принципи підрахунку балів та тлумачення

[ ] SS-1 (15) Тлумачте будь-які відмінності в групових результатах спираючись на всю доступну актуальну інформацію.

[ ] SSI-2 (16) Зіставляйте бали популяцій лише коли встановлено рівень інваріантності в звітній шкалі.

Принципи документування

[ ] Doc-1 (17) Надавайте технічну документацію з будь-якими змінами, враховуючи докази еквівалентності, коли тест адаптується під нову популяцію.

[ ] Doc-2 (18) Забезпечте користувачів тесту документами, що сприятимуть успішному застосуванню адаптованого інструмента у контексті нової культури.

ДОДАТОК В. Глосарій термінів

Альфа (коефіцієнт альфа, альфа Кронбаха). Коефіцієнт надійності тесту, завдання котрого, за припущенням, мають вимірювати ту саму властивість та мати рівний розподіл (тому для особливих випадків існує Омега – див. нижче). Зазвичай є нижнім порогом надійності.

Версія тесту оригінальною мовою. Мова, якою було написано оригінальний тест.

Версія тесту цільовою мовою (перекладена версія). Версія тесту тією мову, якою здійснюється переклад чи адаптація. Наприклад, якщо тест перекладався з англійської іспанською, англійська версія буде називатися «оригінальною версією», а іспанська – «цільовою».

Диференціальне функціонування завдань. Клас статистичних процедур, за допомогою яких можна встановити, чи функціонує завдання більш-менш однаково в двох різних групах. Порівняння у виповненні тесту спершу здійснюється через зіставлення учасників на основі вимірюваної тестом риси. Коли можна спостерігати відмінності, це говорить про те, що завдання містить можливу похибку. Робиться все необхідне, щоби пояснити респондентам груп, зіставлених на основі вимірюваної завданням властивості, про розбіжності у виповненні тесту та чим вони зумовлені.

Дімензіональність (розмірність) тесту. Мова йде про кількість вимірів або факторів, що вимірюються тестом. Часто цей аналіз проводиться завдяки одній зі статистичних процедур, включаючи власні значення чи моделювання структурними рівняннями.

Екзаменовані. Використовується у сфері тестування як синонім до слів «учасники тесту», «протестовані», «респонденти», або «студенти» (в тому випадку, якщо мова йде про тести досягнень).

Значення коефіцієнта дельта. Величини коефіцієнта дельта є просто нелінійно перетвореними p-значеннями, що застосовуються до завдань, які піддаються дихотомічній моделі обробки. Значення дельта завдання є нормальним відхиленням, що відповідає площі під нормальним розподілом (середнє значення=0.0, стандартизація даних=1.0)., де площа під нормальним розподілом рівна долі респондентів, що правильно відповідають на завдання. То ж, якщо p=.84, тоді величина дельта завдання буде -1.0. Це перетворення проводиться з переконанням, що значення дельта з більшою вірогідністю присутнє на шкалі з рівними інтервалами, ніж p-значення.

Конфірматорний факторний аналіз. Висувається гіпотеза з приводу структури тесту, після чого проводяться аналізи для отримання тестової структури із матриці кореляції завдань у тесті. Проводиться статистичний тест для виявлення того, чи подібні між собою передбачувана та вирахувана структури настільки, що гіпотезу про їх рівнозначність не може бути відкинуто.

Локалізація. Популярний термін у сфері тестування, що означає процес пристосування тесту до нового мовного та культурного середовища. Терміном-синонімом є переклад/адаптація.

Метод зворотнього перекладу. Сутність даного методу полягає в тому, що вже перекладений з оригінальної мови тест перекладається іншим перекладачем або групою перекладачів у зворотньому напрямку. Після цього результат зворотнього перекладу порівнюється з оригінальним тестом: якщо ступінь їхньої схожості досить велика, можна зробити припущення, що версія тесту цільовою мовою є цілком прийнятною.

Метод логістичної регресії для виявлення диференціального функціонування завдань. Дана статистична процедура є ще одним способом провести аналіз DIF. Логістична крива підганяється під дані з виповнення тесту кожною групою, після чого дві логістичні криві – одна на кожну групу-носія мов, порівнюються статистичним чином.

Метод Мантеля-Хензеля для виявлення диференціального функціонування завдань. Статистична процедура для порівняння виповнення тестового завдання двома групами респондентів. Порівняння проводяться серед учасників кожної групи, що зіставлялись на основі вимірюваного тестом конструкту чи риси.

Метод подвійного перекладу та узгодження. Незалежний перекладач чи група експертів виявляє та вирішує будь-які розходження між альтернативними версіями прямого перекладу, поєднуючи їх у єдину версію.

Моделювання структурними рівняннями. Комплекс складних статистичних моделей, що використовуються для визначення внутрішньої структури тесту або набору тестів. Ці моделі часто використовуються для дослідження причинних взаємозв'язків між змінними.

Одночасна розробка тесту. Розробка версій тесту оригінальною та цільовою (цільовими) мовами, що проводиться із використанням стандартизованих процедур контролю якості перекладу. Великомасштабні міжнародні проекти все більше вдаються до одночасної розробки, щоби уникнути проблем із складнощами перекладу всіма передбачуваними мовами версії, розробленої одномовною.

Оцінка факторної структури. Факторний аналіз є статистичною процедурою, що застосовується, наприклад, із матрицею кореляції, виробленої завдяки внутрішнім кореляціям поміж ряду завдань тесту (або набору тестів). Мета полягає в спробі пояснити внутрішні кореляції серед тестових завдань з точку зору невеликої кількості факторів, котрі, як вважається, вимірюються тестом (або тестами). Наприклад, за допомогою математичного тесту, факторний аналіз може виявити, що завдання поділяються на три кластери – завдання на здібність до рахування, визначення понять та вирішення задач. Тоді можна сказати, що тест вимірює три фактори – числення, математичні поняття та вирішення математичних задач.

Омега (коефіцієнт омега, омега МакДональда). Коефіцієнт надійності тесту, завдання якого мають вимірювати ту саму властивість (відповідає моделі загального фактору). Є більш загальноприйнятним ніж альфа.

Прирівнювання тестових балів. Статистична процедура, що виповнюється задля зв'язування оцінок, отриманих після проведення двох тестів, що вимірюють той самий конструкт, але не є суто паралельними.

Теорія відповідей на тестові завдання. Клас статистичних моделей для зв'язування відповідей на завдання з властивістю чи набором рис, які вимірюються завданнями тесту. Певні моделі IRT можуть вправитися як з дихотомними даними, так і з політомними. Дихотомні дані можуть бути отримані з обробки балів за завдання множинного вибору чи завдання з вибором відповідей типу «вірно/невірно» в шкалі особистісних характеристик. Політомні дані можуть бути отримані з обробки балів за випробування робочих характеристик, есе, тести навчальної успішності, або виповнення завдань шкали Лікерта.

Формула К'юдера-Річардсона (KR-20). Коефіцієнт надійності тесту, сформований на основі дихотомічних завдань, які мають вимірювати одну спільну властивість та мати рівний розподіл.

PISA (англ. Programme for International Student Assessment). Міжнародна оцінка навчальної успішності, що спонсорується Організацією економічного співробітництва та розвитку (OECD), та в якій приймають участь більш ніж 40 країн.

TIMSS (англ. Trends in International Mathematics and Science Studies). Міжнародне порівняльне дослідження якості та рівня природничо-математичної освіти учнів 4-х, 8-х та 12-х класів, що спонсорується Міжнародною асоціацією з оцінки навчальної успішності IEA.

WDMS (англ. Weighted Multidimensional Scaling). Зважене багатовимірне шкалювання – ще один статистичний метод дослідження розмірності (дімензіональності) тесту.