**МОДЕЛЬ РЕЦЕНЗУВАННЯ ЄФПА**

**ДЛЯ ОПИСУ ТА ОЦІНКИ ПСИХОЛОГІЧНИХ ТА НАВЧАЛЬНИХ ТЕСТІВ**

***ФОРМУЛЯР РЕЦЕНЗІЇ ТЕСТУ ТА НОТАТКИ ДЛЯ РЕЦЕНЗЕНТІВ***

**ВЕРСІЯ 4.2.6**

Версія 4.2.6 є результатом ґрунтовного перегляду Версії 3-42 (2008)

спеціальної групи з комітету оцінювання ЄФПА, до якої входять:

Арне Еверс (голова комітету; Нідерланди)

Кармен Хагемайстер (Німеччина)

Андреас Хьостмелінген (Норвегія)

Патріція Ліндлей (Велика Британія)

Хосе Муньїс (Іспанія)

Андерс Шьоберг (Швеція)

***Схвалено Генеральною Асамблеєю ЄФПА, 13.07.2013***

**© EFPA** За вимогою ЄФПА, користувачі даного документу та його змісту мають посилатися на дане джерело, використовуючи наступний текст:

*“The EFPA Test Review Criteria were largely modelled on the form and content of the British Psychological Society's (BPS) test review criteria and criteria developed by the Dutch Committee on Tests and Testing (COTAN) of the Dutch Association of Psychologists (NIP). EFPA is grateful to the BPS and the NIP for permission to build on their criteria in developing the European model. All intellectual property rights in the original BPS and NIP criteria are acknowledged and remain with those bodies.”*

**ЗМІСТ**

**1. Вступ**

**ЧАСТИНА 1: ОПИС ІНСТРУМЕНТА**

**2. Загальний опис**

**3. Класифікація**

**4. Вимірювання і підрахунок**

**5. Комп’ютерні звіти**

**6. Умови постачання та витрати**

**ЧАСТИНА 2: ОЦІНКА ІНСТРУМЕНТА**

**7. Якість пояснення теоретичної основи та поданої інформації**

**7.1 Якість пояснення теоретичної основи**

**7.2 Адекватність доступної користувачеві документації**

**7.3 Якість методичних інструкцій для користувача**

**8. Якість тестових матеріалів**

**8.1 Якість матеріалів тестів бланкового типу**

**8.2 Якість матеріалів комп’ютерних та Інтернет-тестів**

**9. Норми**

**9.1 Нормативно-орієнтована інтерпретація**

**9.2 Критеріально-орієнтована інтерпретація**

**10. Надійність**

**11. Валідність**

**11.1 Конструктна валідність**

**11.2 Критеріальна валідність**

**11.3 Загальна валідність**

**12. Якість комп’ютерних звітів**

**13. Остаточне оцінювання**

**ЧАСТИНА 3: БІБЛІОГРАФІЯ**

**ДОДАТОК. Пам’ятка ключових пунктів для коментування у випадку, коли інструмент було перекладено та/або адаптовано з нелокального контексту**

**1. Вступ**

Основну мету Моделі рецензування тестів ЄФПА становить надання опису і детальної та ретельної оцінки психологічних оціночних випробувань, шкал та опитувальників у сферах праці, освіти, охорони здоров'я та інших. Цю інформацію буде зроблено доступною для користувачів тестів та експертів в сфері тестування, щоби сприяти покращенню тестів та самого процесу тестування, а також допомагати користувачам та експертам приймати вірні оціночні рішення. Модель рецензування тестів ЄФПА є частиною її інформаційної стратегії, мета якої полягає в тому, щоби надати оцінку всієї необхідної технічної інформації про тести для поширення їх використання (Evers et al., 2012; Muñiz & Bartram, 2007). Згідно зі *“Стандартами психологічного та освітнього тестування”*, терміном “тест” охоплюється будь-який “оцінювальний засіб або процедура під час якої отримується вибірка поведінки об’єкту дослідження у певному полі діяльності, після чого її оцінюють, та підраховують результати, використовуючи стандартизований процес” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 3). Таким чином, ця модель рецензування застосовується до всіх інструментів, що підпадають під дану дефініцію, незалежно від того, як вони називаються -- шкалами, опитувальниками, проективною методикою і т.д.

Оригінальна версія моделі рецензування тестів ЄФПА була створена на основі великої кількості джерел, серед яких: *Оцінювальна форма рецензування тестів* від Британського товариства психологів (розроблена Newland Park Associates Limited, NPAL та привласнена Керівним комітетом з тестових стандартів БТП), *Іспанський опитувальник оцінювання психометричних тестів* (розроблений Іспанською асоціацією психологів), та *Рейтингова система якості тестів* (розроблена Голландським комітетом з тестів та тестування Голландської асоціації психологів). Було дозволено адаптувати значну частину змісту проформи рецензування, яка була створена Newland Park Associates Ltd  у 1989 р. для навчальних засобів з Великої Британії (див. Bartram, Lindley & Foster, 1990). Згодом вона була використана і розроблена в серії рецензій БТП для інструментів, що використовуються в оцінюваннях при наймі (наприклад, Bartram, Lindley, & Foster, 1992; Lindley et al., 2001). Перша версія моделі рецензування ЄФПА була скомпільована та відредагована Дейвом Бартрамом (Bartram, 2002a, 2002b) після симпозіуму ЄФПА в березні 2000 і ряду консультацій, що відбулись пізніше. Значні корективи були внесені Патріцією Ліндлей, Дейвом Бартрамом, та Наталі Кеннеді для використання в системі рецензування БТП (Lindley et al, 2004). Це в подальшому було адаптовано ЄФПА в 2005 (Lindley et al., 2005), з невеликими поправками в 2008 (Lindley et al., 2008). Поточну версію моделі підготовила спеціальна група з комітету оцінювання ЄФПА, до якої входили: Арне Еверс (голова комітету; Нідерланди), Кармен Хагемайстер (Німеччина), Андреас Хьостмелінген (Норвегія), Патріція Ліндлей (Велика Британія), Хосе Муньїс (Іспанія) та Андерс Шьоберг (Швеція). В цю версію було включено нотатки та чек-лист для перекладених та адаптованих тестів від Патріції Ліндлей та редакторів-консультантів рецензування тестів Великої Британії (Lindley, 2009). Тексти деяких значно поновлених та відредагованих фрагментів базуються на переглянутій версії Голландської рейтингової системи для якості тесту (Evers, Lucassen, Meijer, & Sijtsma, 2010;Evers, Sijtsma, Lucassen, & Meijer, 2010).

Модель рецензування тестів ЄФПА поділяється на три основні частини. В першій частині (Опис інструменту) наданий детальний опис усіх властивостей тесту. В другій частині (Оцінка інструменту) оцінюються фундаментальні якості тесту: матеріали, норми, надійність, валідність, та комп’ютерні звіти, включно з загальною фінальною оцінкою. У третій частині (Бібліографія) містяться використані джерела.

Не менш за саму модель важливим є належне її практичне застосування. Поточна версія моделі призначена для використання двома незалежними рецензентами у процесі експертної оцінки, подібно до звичайної оцінки наукових письмових робіт та проектів. Редактор-консультант повинен переглянути рецензії, та залучити третього рецензента в тому випадку, якщо між двома рецензіями будуть виявлені значні розбіжності. Деякі відхилення в процедурі можливі, доки гарантована компетентність та незалежність рецензентів і редактора-консультанта. ЄФПА рекомендує направляти оцінки рецензій кваліфікованим практикуючим користувачам тестів, але вони також можуть становити інтерес для вчених, авторів тесту та експертів з психометричного та психологічного тестування.

Іншим важливим питанням є публікація результатів оцінювання тесту. Результати мають бути доступними всім експертам та користувачам (платно чи безкоштовно). Вдалим варіантом буде опублікувати результати на веб-сайті Національної психологічної асоціації, але вони також можуть бути опубліковані третіми сторонами чи іншими засобами масової інформації – у журналах чи книгах.

Первісною метою розширення доступності цієї моделі є підтримка узгодженості рецензувальних процедур та критеріїв в межах Європи. Однак узгодженість є лише однією з цілей моделі, інша мета -- запропонувати систему рецензування тестів країнам, які не мають своїх власних процедур рецензування. Зрозуміло, що питання місцевого значення неминуче потягнуть за собою зміни в Моделі рецензування тестів ЄФПА або в процедурах рецензування, коли Модель почне використовуватися в інших країнах. Саме тому Модель назвали *Моделлю* -- щоби підкреслити можливість здійснення локальних адаптацій для гарантії ефективного задоволення місцевих потреб.

Заохочуються будь-які коментарі та зауваження до моделі рецензування тестів ЄФПА, адже вони можуть стати у нагоді при покращенні та поясненні процедур.

**ЧАСТИНА 1: ОПИС ІНСТРУМЕНТУ**

**2. Загальний опис**

Цей сегмент формуляру має надавати загальну інформацію про те, що це за інструмент і де його можна отримати. Він має містити назву інструменту, видавця та дистриб’ютора, автора, дату оригінальної публікації та версії, що рецензується.

Відповіді на пункти 2.1.1. - 2.7.3 мають бути простими та чіткими. Вони несуть у собі фактичну інформацію. Однак при заповненні пунктів щодо областей інформаційного змісту може потребуватися самостійне осмислення.

|  |  |  |
| --- | --- | --- |
|  | Рецензент[[1]](#footnote-1) |  |
|  | Дата поточного рецензування |  |
|  | Дата попереднього рецензування (за потреби)[[2]](#footnote-2) |  |
| 2.1.1 | Назва інструменту (локальна версія) |  |
| 2.1.2 | Коротка назва тесту (за потреби) |  |
| 2.2 | Оригінальна назва тесту (якщо локальна версія адаптована) |  |
| 2.3 | Автори оригінального тесту |  |
| 2.4 | Автори локальної адаптації |  |
| 2.5 | Локальний дистриб’ютор/видавець тесту |  |
| 2.6 | Видавець оригінальної версії тесту (якщо відрізняється від поточного видавця/дистриб’ютора) |  |
| 2.7.1 | Дата публікації поточної ревізії/редакції |  |
| 2.7.2 | Дата публікації адаптації для локального використання |  |
| 2.7.3 | Дата публікації оригінального тесту |  |

|  |
| --- |
| **Загальний опис інструменту**  Короткий незалежний безоціночний опис (200-600 слів)  Тут треба надати стислий безоціночний опис інструменту. Опис повинен надавати чітке уявлення про те, що представляє собою інструмент – його зміст, які фактори призначений вимірювати і т.д. Він має бути написаний в нейтральному стилі, та повідомляти про сутність інструменту, які фактори він вимірює, яка його мета застосування, наявність груп норм та їх тип, незвичайні характеристики та історичне підґрунтя. Опис повинен бути досить невеликим (200-300 слів). Однак для деяких більш складних багатофакторних інструментів він може бути довшим (300-600 слів). Він має бути написаним так, щоби слугувати як самостійний опис інструменту. Як наслідок, він може повторювати більш специфічну інформацію, що надається у відповідях в сегментах 2–6. Він має в загальних рисах описувати всі доступні версії інструменту, посилання на котрі зустрінуться на наступних сторінках.  На даний пункт слід відповідати спираючись на інформацію, надану видавцем, котру, в свою чергу, рецензент має перевірити на точність. |
|  |

**3. Класифікація**

|  |  |  |
| --- | --- | --- |
| 3.1 | **Області інформаційного змісту** *(помітьте всі підходящі варіанти)*  Ви маєте ідентифікувати всі передбачені видавцем області інформаційного змісту. Якщо ж вони чітко не визначені, треба відповідати спираючись на інформацію, подану в керівництві (вибірка стандартизації, методи і області застосування, валідність, і т.д.). | * Здібності – Загальні * Здібності – Мануальні навички/дрібна моторика * Здібності – Технічні * Здібності – Здатність до навчання/пам'ять * Здібності – Розуміння абстракцій/ невербальний інтелект/ індуктивне мислення * Здібності – Математичні * Здібності – Швидкість сприйняття * Здібності – Сенсорно-рухові * Здібності – Орієнтування в просторі, зорові здібності * Здібності – Вербальні * Увага/концентрація * Переконання * Когнітивний стиль сприйняття * Розлади та патології * Функціонування сім’ї * Функціонування групи * Інтереси * Мотивація * Функціонування організації, робоча обстановка * Особистість – Особистісні риси * Особистість – Типи особистості * Особистість – Психічний стан * Якість життя * Оцінка успішності (тест на ефективність засвоєння матеріалу) * Шкільна чи навчальна обстановка * Ситуаційні рішення * Стрес/ емоційне вигорання * Результати терапії * Цінності * Благополуччя * Інше (впишіть): |
| 3.2 | **Передбачувані або основні сфери застосування** *(помітьте всі підходящі варіанти)*  Ви маєте ідентифікувати передбачені видавцем сфери застосування інструменту. Якщо ж вони чітко не визначені, треба відповідати спираючись на інформацію, подану в керівництві (вибірка стандартизації, методи і області застосування, валідність, і т.д.). | * Медицина * Профорієнтація * Освіта * Судова експертиза * Загальний стан здоров'я, рівень життя та благополуччя * Неврологія * Спорт та дозвілля * Сфера праці та найму * Інше (впишіть): |
| 3.3 | **Опис груп людей, для яких тест призначений**  На цей пункт слід давати відповідь спираючись на інформацію від видавця.  Для деяких тестів групи можуть бути обширнішими (напр., група повнолітніх), для інших — більш специфічними (напр., робітники середньої і важкої фізичної праці, або хлопчики від 10 до 14 років).  В даному розділі мають бути зазначені лише чітко встановлені групи. Якщо вибір цих груп здається недоцільним, про це треба зазначити в тій частині рецензії, де надається оцінка інструмента. |  |
| 3.4 | **Число факторів та стислий опис змінних, що вимірюються інструментом**  На цей пункт слід давати відповідь спираючись на інформацію від видавця. Визначте кількість факторів (якщо >1) та надайте короткий опис кожного з них, якщо його сутність не зрозуміла з назви. Рецензія інструменту має включати в себе розгляд інших похідних оцінок, які зазвичай використовуються при застосуванні даного інструменту та описані в стандартній документації – наприклад, первинні бали якостей, а також «Велика п’ятірка» вторинних балів якостей для комплексного особистісного тесту, або субтесту, фактору, і підсумкових балів тесту розумових здібностей. |  |
|  | **Форма відповіді**    На цей пункт слід давати відповідь спираючись на інформацію від видавця.  Якщо потребується будь-яке певне встаткування (котре не зазначено у списку, наприклад, цифровий реєстратор), воно має бути зазначене тут. Крім того, треба надати опис будь-яких специфічних умов тестування. “Стандартні умови тестування” вважаються придатними для тестування під наглядом інспектора. Такими умовами є тиха, добре освітлена та провітрена кімната з достатньою кількістю столів та стільців для інспекторів та учасників тестування. | * Усне інтерв’ю * Письмовий тест * Ручна (фізична) діяльність * Безпосереднє спостереження * Комп’ютеризований тест * Інше (впишіть): |
| 3.6 | **Вимоги до учасника тесту**  На цей пункт слід давати відповідь спираючись на інформацію від видавця. Які здібності та навички необхідні учасникові тесту для проходження тестування і отримання справедливої інтерпретації результатів? Зазвичай очевидно, коли повна нестача необхідних умов робить проходження тесту неможливим (наприклад, якщо учаснику з вадами зору дати письмовий тест), але всі необхідні вимоги мають бути зазначені та класифіковані таким чином:   * “Не має значення” -- мається на увазі, що певні вміння або здібності взагалі непотрібні — наприклад, мануальні здібності для усних відповідей на запитання. * “Надана необхідна інформація” значить, що присутні певні обмеження. * “Інформація відсутня” означає, що можуть бути присутні обмеження для учасників без певних навичок чи здібностей (що відомо теоретично чи з емпіричних результатів), але це чітко не зазначено чи не зрозуміло з інформації наданої видавцем, наприклад, якщо тест використовує мову яка не є рідною мовою учасника тесту. | Мануальні здібності *(один варіант)*   * не має значення * надана необхідна інформація * інформація відсутня   Домінування однієї з рук *(один варіант)*   * не має значення * надана необхідна інформація * інформація відсутня   Зір *(один варіант)*   * не має значення * надана необхідна інформація * інформація відсутня   Слух *(один варіант)*   * не має значення * надана необхідна інформація * інформація відсутня   Володіння мовою тесту (розуміння та розмовні навички) *(один варіант)*   * не має значення * надана необхідна інформація * інформація відсутня   Навички читання *(один варіант)*   * не має значення * надана необхідна інформація * інформація відсутня   Письмові навички *(один варіант)*   * не має значення * надана необхідна інформація * інформація відсутня |
| 3.7 | **Формат відповідей** *(помітьте один варіант)*  На цей пункт слід давати відповідь спираючись на інформацію від видавця.  Слід розрізняти два типи формату множинного вибору. Перший тип — це тест, в якому треба обрати одну вірну відповідь серед певної кількості варіантів (наприклад, як у тестах на абстрактне мислення). Другий тип – опитувальники, що не містять певної вірної відповіді. Даний формат вимагає обрати між двома чи більше варіантами, що відносяться до різних факторів (наприклад, факторів опитувача професійних інтересів, або особистісного тесту). Такий формат також називається «багатофакторним», тому що варіанти відповідей відносяться до різних факторів. У тестах цього формату може потребуватися ранжувати певні твердження, або обирати відповіді на кшталт «більш схоже на мене» чи «менш схоже на мене». Результати тесту такого формату можуть підраховуватись за допомогою іпсативних шкал (дивіться пункт 3.8).  У шкалі оцінок Лікерта учасник тесту також має обирати серед певної кількості варіантів, але істотна відмінність від формату множинного вибору полягає в тому, що шкали – однофакторні (варіанти відповідей простягаються, наприклад, від «ніколи» до «завжди», чи від «маловірогідно» до «дуже вірогідно»). Інша відмінність – учасник тесту не має обирати між варіантами з різних факторів. Також, шкала має позначатися як шкала Лікерта в тому випадку, коли наявні лише два варіанти для одного фактору (наприклад, так/ні, чи завжди/ніколи). | * Множинного вибору (тест здібностей, або тест з вибором правильних/хибних відповідей)   Кількість варіантів: …   * Множинного вибору (багатофакторний)   Кількість варіантів: …   * Шкала оцінок Лікерта   Кількість варіантів: …   * Відкритий формат * Інший (опишіть): |
| 3.8 | **Іпсативність**  Як зазначається у пункті 3.7, результати тестів множинного вибору багатофакторного формату *можуть* підраховуватись за допомогою іпсативних шкал. Характерною рисою іпсативних шкал є те, що показник кожної шкали або фактору обмежений показниками інших факторів. В повністю іпсативних інструментах сума шкальних балів є незмінною для кожної людини.  Інші процедури підрахунку можуть призводити до іпсативності (наприклад, віднімання загального середнього кожної людини від її шкальованих балів). | * Так, результатом тесту багатофакторного формату множинного вибору є іпсативні бали (частково чи повністю) * Так, результатом тесту іншого формату є іпсативні бали (частково чи повністю) * Ні, іпсативні бали **не є** результатом тесту багатофакторного формату множинного вибору * N/A |
| 3.9 | **Загальна кількість питань тесту та кількість питань на одну шкалу чи субтест**  На цей пункт слід давати відповідь спираючись на інформацію від видавця.  Якщо інструмент має кілька факторів або субтестів, треба зазначити загальну кількість питань чи завдань, та кількість питань на кожний фактор чи субтест. Якщо питання охоплюють більш ніж один фактор чи субтест, це повинно бути задокументовано. |  |
| 3.10 | **Передбачуваний спосіб використання (при яких умовах інструмент було розроблено та затверджено)** *(помітьте всі підходящі варіанти)*  Цей пункт важливий тому, що визначає, чи був інструмент створений для того, щоби використовуватись в умовах відсутності нагляду експертів або контролю адміністрації. Зверніть увагу, що способи використання можуть варіюватись в залежності від версії інструменту. На цей пункт слід давати відповідь спираючись на перевірену інформацію від видавця.  Примітка: чотири способи проведення описані в *Міжнародному Керівництві з* *проведення комп’ютерних та інтернет-тестувань* (International Test Commission, 2005, pp. 5-6). | * *Відкритий спосіб проведення*: Без прямого нагляду експертів за сеансом тестування, а отже без необхідності автентифікації особи учасника тесту. Прикладом даного способу проведення можуть слугувати Інтернет-тести, що не вимагають реєстрації для проходження. * *Проведення під контролем*: Без прямого нагляду експертів за сеансом тестування, при цьому тест доступний тільки відомим учасникам тесту. Інтернет-тести вимагатимуть від учасника тесту зареєструватися та отримати логін і пароль. Частіш за все такі тести призначені для разового проходження. * *Спосіб проведення під наглядом*: З прямим наглядом експертів за умовами проходження тесту. У цьому способі особа учасника тесту може бути ідентифікована. У випадку з інтернет-тестуванням може потребуватися, щоби інспектор залогінив учасника та підтвердив, що тест проводився та складався належним чином. * *Керований спосіб проведення:* З високим рівнем живого людського нагляду та контролю за умовами проходження тесту. У контексті комп’ютерного тестування цього може бути досягнуто із використанням спеціалізованих центрів тестування, де високий рівень контролю над доступом, безпекою, кваліфікацією членів адміністрації, якістю, а також технічними умовами та встаткуванням. |

|  |  |  |
| --- | --- | --- |
| 3.11 | **Спосіб проведення** *(помітьте всі підходящі варіанти)*  На цей пункт слід давати відповідь спираючись на інформацію від видавця.  Якщо необхідні будь-які технічні пристрої чи інші елементи обладнання (окрім тих, що зазначені у списку, наприклад, цифровий реєстратор), вони повинні бути описані тут. Будь-які специфічні умови тестування також мають бути описані тут. “Стандартні умови проведення” передбачені для тестування під наглядом чи контролем. Серед таких умов: тиха, добре освітлена та провітрена кімната, в якій достатньо простору, столів та стільців для інспекторів та учасників тесту. | * Інтерактивне індивідуальне проведення * Групове проведення під наглядом * Комп'ютеризоване тестування у вигляді локально встановленої програми – під наглядом/контролем * Комп'ютеризоване тестування у вигляді веб-програми – під наглядом/контролем * Комп'ютеризоване тестування у вигляді локально встановленої програми – без нагляду/самооцінка * Комп'ютеризоване тестування у вигляді веб-програми – без нагляду/самооцінка * Інше (впишіть): |
| 3.12 | **Скільки часу потребується на використання** **інструменту** *(зазначте час для кожного способу проведення)*  На цей пункт слід давати відповідь спираючись на інформацію від видавця. Відповідь на це запитання можна розбити на декілька компонентів. У більшості випадків можливо лише приблизно оцінити кількість часу, ніж надати точні числа. Головне – надати потенційному користувачеві гарне уявлення про те, скільки часу може потребуватися на використання цього інструменту. Не треба зазначати час, необхідний для ознайомлення із інструментом. Зробіть припущення, що користувач достатньо кваліфікований та досвідчений.   * Час на підготовку (скільки часу треба інспектору для підготовки та роздачі матеріалів для тестування; час на отримання доступу та введення логіну і паролю для проведення тесту онлайн). * Час на проведення тестування: включає в себе час, потрібний для того, щоби завершити всі завдання, а також для надання інструкцій, розгляду зразків завдань, та брифінгу в кінці тестування. * Підрахунок результатів: час на отримання первинних балів. В багатьох випадках цей процес автоматизовано. * Аналіз: скільки часу займає подальша робота над первинними балами, виведення інших оцінок на їх основі, та обґрунтоване і вичерпне тлумачення (передбачається, що ви знайомі з інструментом). Знову ж таки, цей процес може бути автоматизовано. * Зворотній зв'язок: час, необхідній для надання зворотного зв'язку учасникові тесту та іншим зацікавленим сторонам.   Визнається, що час на останні два компоненти може значно варіюватися в залежності від контексту, в якому інструмент використовується. В будь-якому випадку, які завгодно коментарі та зауваження стануть у нагоді. | Підготовка:  Проведення:  Підрахунок:  Аналіз:  Зворотній зв'язок: |
| 3.13 | **Зазначте, чи наявні інші версії інструменту, і які саме є об’єктом рецензування**  Повідомте, чи доступні альтернативні версії (справжні або псевдо-паралельні форми, скорочені версії, комп’ютерні версії, і т.д.) інструменту, та опишіть їх застосування для різних груп людей. В деяких випадках різні версії інструменту вважаються рівноцінними, тобто взаємозамінними альтернативними формами. В інших випадках різні версії призначаються для різних груп (наприклад, версії для дітей і для дорослих). Якщо версій більше однієї, зазначте, чи є вони рівноцінними/альтернативними, або ж вони призначені для виповнення різних функцій – як, наприклад, скорочені та довші версії; іпсативна та нормативна версії. Також опишіть, чи можуть окремі частини використовуватися замість цілого інструменту. Якщо наявні комп’ютеризовані версії, стисло опишіть вимоги до програмного забезпечення та апаратних засобів. Зверніть увагу на те, що якщо доступні автономні комп’ютерні тести та онлайн пакети, це також має бути зазначено. |  |

**4. Вимірювання і підрахунок**

|  |  |  |
| --- | --- | --- |
| 4.1 | **Процедура підрахунку результатів тесту** *(помітьте всі підходящі варіанти)*  Цей пункт слід заповнювати, посилаючись на інформацію від видавця, різноманітні керівництва та документи.  Сервісний центр – це послуги, що надаються поставником, чи його агентом, для підрахунку та тлумачення. Дані послуги є факультативними. Якщо провести підрахунок або тлумачення можна лише за допомогою сервісного центру, про це необхідно зазначити в рецензії, та написати про витрати в пункті поточних витрат. | * Комп’ютерний підрахунок з прямим введенням відповідей учасником тесту * Комп’ютерний підрахунок з введенням відповідей з бланку за допомогою оптичного зчитувача міток * Комп’ютерний підрахунок з введенням відповідей вручну з бланку * Простий підрахунок вручну з використанням ключів – вимагаються лише канцелярські навички * Складний підрахунок вручну – вимагається проходження навчання підрахунку результатів інструменту * Сервісний центр -- наприклад, підрахунок результатів компанією, що продає інструмент * Інше (опишіть): |
| 4.2 | **Показники**  Цей пункт слід заповнювати, посилаючись на інформацію від видавця, різноманітні керівництва та документи.  Стислий опис системи підрахунку балів для отримання загальних та проміжних показників, поправка на випадкове вгадування, допоміжні засоби якісного тлумачення, і т.д. |  |
| 4.3 | **Шкали** *(помітьте всі підходящі варіанти)*  Цей пункт слід заповнювати, посилаючись на інформацію від видавця, різноманітні керівництва та документи. | *Показники, основані на процентілях*   * Центілі * П’ятиступінчаста класифікація: 10:20:40:20:10 центільні поділки * Децилі * Інші (впишіть):   *Стандартні показники*   * Z-показники * Відхилення показників IQ (наприклад, середнє значення 100, стандартизація даних за Векслером = 15, або 16 за Станфорд-Бінетом) * Прийомна екзаменаційна комісія коледжу (наприклад, середнє значення тесту з перевірки здібностей = 500, стандартизація даних = 100) * Стени * Стенайни * Т-показники * Інше (впишіть): * Критичні показники, прогностичні таблиці, та інші показники, орієнтовані на певні рішення * Лише первинні бали * Інше (впишіть): |
| 4.4 | **Перетворення стандартних оцінок** | * Нормалізоване -- стандартні показники отримуються з використанням довідкової таблиці з нормалізації * Не нормалізоване -- стандартні показники отримуються з використанням лінійного перетворення * N/A |

**5. Комп'ютерні звіти**

**Зверніть увагу, що даний сегмент цілком описовий. Оцінки звітів мають бути надані в частині рецензії “Оцінка інструменту”.**

У випадку наявності декількох чи більше генерованих звітів, заповніть пункти 5.2 - 5.13 для кожного з них чи незалежної секції звіту (по мірі необхідності копіюйте сторінки). Дана система класифікації може бути використана для описання двох генерованих звітів, наприклад, Звіт №1 може призначатися для учасника тесту, або для інших ненавчених користувачів, а Звіт №2 -- для навченого користувача, що достатньо компетентний для використання інструменту та обізнаний щодо того як його тлумачити.

|  |  |  |
| --- | --- | --- |
| 5.1 | **Чи наявні комп’ютерні звіти?**  Якщо відповідь до 5.1 – «так», то потребується використати нижченаведену класифікацію для систематизації типів наявних звітів. У багатьох інструментів буде немалий ряд доступних звітів. Будь ласка, заповніть окрему форму для кожного з них. | * Так (заповніть подальші пункти) * Ні (переходьте до пункту 6.1) |
| 5.2 | **Назва чи опис звіту**  *(дивіться вступ до цього сегменту)* |  |
| 5.3 | **Засоби передачі інформації** *(помітьте всі підходящі варіанти)*  Звіт може складатися лише з одного тексту, чи містити також зображення показників у вигляді таблиці чи графіка (наприклад, шкала стенів). Там, де є і текст, і графічні дані, вони можуть бути викладені як паралельно один одному, так і бути пов’язаними між собою таким чином, щоби проглядався зв'язок між твердженнями з тексту та показниками. | * Лише текст * Не пов’язані між собою текст та графічні дані * Пов’язані між собою текст та графічні дані * Лише графічні дані |
| 5.4 | **Рівень складності** *(помітьте один варіант)*  Деякі звіти є дуже простими, наприклад такі, де замість текстового блоку викладені Стени з послідовним описом шкал. Інші є більш складними, включаючи в себе блоки тексту, які пов’язані з паттернами та співвідношенням шкальних балів, та в яких розглядаються ефекти взаємодії факторів/шкал. | * Простий (наприклад, перелік пунктів, в кожному з яких надається опис шкали) * Середній (суміш простих описів зі складнішим матеріалом) * Складний (описує паттерни та співвідношення шкальованих балів, а також взаємодію шкал) |
| 5.5 | **Структура звіту** *(помітьте один варіант)*  Структура пов’язана з рівнем складності. | Структура, що базується на:   * Шкалах – коли звіт будується навколо індивідуальних шкал * Факторах – коли звіт будується навколо факторів вищого порядку, як, наприклад, «Велика п’ятірка» для вимірювання індивідуальних якостей * Конструктах – коли звіт будується навколо одного чи більше наборів конструктів (наприклад, в контексті праці можуть бути такі конструкти як типи колективу, стиль керівництва, або стійкість до стресів; в контексті медицини це можуть бути різні психопатологічні синдроми; і т.д.), що прив’язані до оригінальних шкальних оцінок. * Критеріях – коли звіт фокусується на зв’язку з емпіричними результатами (наприклад, шкільна успішність, результати терапії, продуктивність праці, абсентеїзм, і т.д.) * Інше (опишіть): |
| 5.6 | **Чутливість до контексту** *(помітьте один варіант)*  Люди, що пишуть звіти, адаптують мову, формат та контент звіту під потенційного читача, а також приймають до уваги мету оцінювання і контекст, в якому воно буде відбуватися. В контексті праці, звіт, написаний для мети відбору персоналу, відрізнятиметься від звіту для підвищення кваліфікації; звіт для менеджеру середнього віку буде відрізнятися від того, що написаний для молодої людини, починаючої програму навчання на ринку праці. В контексті освіти, звіт, зроблений для оцінки загальної здібності студентів до навчання та функціонування в навчальному середовищі, та звіт, спродукований для оцінки того, чи не має студент порушення здатності до навчання, також будуть різними. Звіт, спрямований на надання пропозицій щодо цілей навчання іншим професіоналам буде відрізнятися від звітів, що інформують батьків про сильні та слабкі сторони їх дітей. В контексті медицини звіт, зроблений задля цілей діагностування, відрізнятиметься від звіту, в якому оцінюється схильність пацієнта до ризикованої поведінки. Звіт для надання зворотного зв'язку пацієнтам буде відрізнятися від звіту, що повідомляє керівництво про те, чи безпечно звільняти пацієнта від примусового лікування. | * Одна версія для всіх контекстів * Версії призначені для певних контекстів; кількість контекстів: … * Контекст визначається користувачем, звіт доступний для редагування |
| 5.7 | **Медико-актуарний звіт** *(помітьте всі підходящі варіанти)*  Більшість систем звітності базуються на клінічній оцінці. Тобто фрагменти тексту напише людина, чи люди, що є компетентними користувачами інструменту, про який йдеться мова. Отже, звіти представлятимуть собою їх особисті тлумачення шкал. Деякі системи містять в собі актуарні звіти, в яких твердження базуються на дослідженнях емпіричної валідизації, що пов’язує шкальні бали, наприклад, з показниками якості роботи, чи з клінічною класифікацією. | * Базується на клінічній оцінці одного експерту * Базується на клінічній оцінці декількох експертів * Базується на емпіричних та актуарних закономірностях |
| 5.8 | **Можливість модифікації** *(помітьте один варіант)*  Висновок (результат) звіту часто є фіксованим. Однак деякі системи виробляють звіти у вигляді файлу, котрий може редагуватися користувачем. Інші можуть надавати інтерактивний доступ онлайн як для кінцевого користувача, так і для учасника тесту. | * Без можливості модифікування (фіксована печатна форма висновку) * Обмежена модифікація (обмежена до певних областей, наприклад, до області професійної біографії) * Необмежена модифікація (наприклад, через доступ до файлу текстового редактору) * Інтерактивний звіт, який надає учасникові тесту можливість вносити коментарі, чи надає дані про обґрунтованість змісту (наприклад, через груповий онлайн доступ до інтерактивного механізму продукування звітів) |
| 5.9 | **Рівень завершеності** *(помітьте один варіант)*  До якої міри готовності система призначена генерувати інтегровані тексти – у готовому для використання вигляді, чи як набір приміток, коментарів, гіпотез, і т.д. | * Топографічної якості * Чорнової якості |
| 5.10 | **Прозорість** *(помітьте один варіант)*  Системи відрізняються в плані відкритості або прозорості для користувача. Відкрита система – це система, де чітко простежується зв’язок між шкальними балами та текстом. Така відкритість можлива лише коли наявні і текст, і показники, а зв’язок між ними цілком очевидний. Інші системи працюють за принципом «чорного ящику», що ускладнює спробу користувача встановити зв'язок між шкальними балами та текстом. | * Відкритий зв'язок між конструктами, показниками та текстом * Скритий зв'язок між конструктами, показниками та текстом * Суміш відкритого та скритого зв’язку між конструктами, показниками та текстом |
| 5.11 | **Характер та стиль** *(помітьте один варіант)*  Системи також відрізняються за наповненістю директивами та гіпотезами. В робочому та організаційному контексті таке твердження як «Пан Ікс дуже сором’язливий -- з нього не вийде успішного торгового агенту» є директивним, тоді як інші твердження оформлені у вигляді гіпотез та орієнтирів: «Згідно з балами пана Ікс по фактору Y, він виявляється досить сором’язливим у порівнянні з референтною групою торгівців. Якщо це справді так, він може зазнати труднощів, працюючи у торгівельному середовищі. Це вимагає подальшого дослідження за його участі». В освітньому контексті директивне твердження може бути наступним: «Результати вказують на те, що в плані математичних здібностей персона Ікс відстає на два роки від середнього показника своїх одноліток», тоді як орієнтаційне твердження може бути таким: «Результати показують, що Ікс легко відволікається на зовнішні подразники під час виповнення завдань. Спостереження поведінки під час тестування це підтверджує. Необхідно прийняти це до уваги під час створення оптимального навчального середовища для Ікс». В медичному контексті директивне твердження буде таким: «Результати тесту вказують на те, що пацієнт має односторонню зорово-просторову агнозію, і тому не спроможний безпечно керувати транспортним засобом», тоді як гіпотетичне твердження може виглядати так: «Результати тесту пані Ікс вказують на можливі проблеми зі встановленням тривалих близьких стосунків. Це потребує подальшого дослідження перед постановкою діагнозу». | * Директивний/інструктивний характер * Орієнтаційний/гіпотетичний характер * Інший (опишіть): |
| 5.12 | **Передбачувані одержувачі** *(помітьте всі підходящі варіанти)*  Звіти зазвичай розробляються, щоби служити інтересам однієї чи більше категорій користувачів. Користувачі загалом поділяються на чотири групи:   1. *Кваліфіковані користувачі тесту.* Це люди, що мають необхідні знання та навички для продукування власних звітів на основі шкальних балів. Вони повинні бути спроможні користуватися звітами, що містять технічну психометричну термінологію та відкриті зв’язки між шкалами та описаннями. Вони також мають вміти адаптувати та модифікувати звіти. 2. *Кваліфіковані користувачі системи.* Не будучи компетентними для генерування власних звітів на базі шкальних показників, люди цієї групи достатньо компетентні для того, щоби користуватися згенерованими системою результатами. Рівень підготовки, необхідний для отримання цієї компетенції може значно варіюватися в залежності від характеру комп’ютерних звітів (наприклад, звіт на основі рис характеру чи звіт на основі компетенцій; складний звіт чи простий) та мети застосування (низькі та високі ставки). 3. *Учасники тесту.* Учасники тестування зазвичай не будуть попередньо ознайомлені ні з інструментом, ні з типом звіту, розробленим системою. Призначені для них звіти мають бути написані такою мовою, що не передбачає знань про психометрію або даний інструмент. 4. *Треті сторони.* Ця група включає інших людей, зацікавлених в інформації представленій в звіті, або тих, хто може отримати копію звіту. Це можуть бути потенційні роботодавці, менеджер чи супервайзер учасника тесту, один з батьків підлітка, що пройшов профорієнтаційную консультацію. Мова звіту, призначеного для даної категорії людей, буде такою ж як у звітах, призначених для учасників тесту. | * Кваліфіковані користувачі тесту * Кваліфіковані користувачі системи * Учасники тесту * Треті сторони |
| 5.13 | **Чи надає дистриб’ютор послуги з модифікації або розробки кастомізованих комп’ютерних звітів?** *(помітьте один варіант)* | * Так * Ні |

**6. Умови постачання та витрати**

Ця таблиця визначає, що буде надавати видавець, кому, за яких умов, та за яку ціну. Також тут будуть зазначені умови поставника щодо того, хто може, а хто не може отримувати матеріали інструменту. Якщо одна з опцій не відповідає умовам поставки, надайте опис належних умов.

|  |  |  |
| --- | --- | --- |
| 6.1 | **Документація від видавця, надана як частина пакету тесту** *(помітьте всі підходящі варіанти)* | * Довідник користувача * Технічний (психометричний) довідник * Додаткова технічна та поточна інформація (наприклад, місцеві норми, місцеві дослідження валідності, і т.д.) * Книги та статті, пов’язані з тематикою тесту |
| 6.2 | **Способи публікації** *(помітьте всі підходящі варіанти)*  Наприклад, технічні довідники можуть поновлюватися та бути доступними для скачування з Інтернету, в той час як довідники користувача надаватися в друкованому варіанті, або на CD/DVD. | * Друк * CD/DVD * Завантажується з Інтернету * Інший (вкажіть): |
| Пункти 6.3. – 6.5. покривають інформацію щодо витрат. Ця інформація може дуже скоро застарівати. Перед публікацією рецензії рекомендується зв’язатися з поставником чи видавцем як можна скоріше, щоби надати актуальну інформацію щодо коштів. | | |
| 6.3.1 | **Пускові витрати**  Ціна на повний набір матеріалів (всі керівництва та інший матеріал, необхідний для проведення якнайменш одного тесту). Зазначте, скільки учасників тесту буде оцінюватися за допомогою матеріалів, придбаних на ці кошти, та чи входять до них матеріали для повторного оцінювання.  У відповіді на цей пункт треба спромогтися встановити «стартові витрати». Це кошти на придбання повного набору довідкових матеріалів, ключів для підрахунку балів і так далі. Вони включають в себе витрати на навчання тільки коли інструмент «закритий», і витрати на специфічну попередню підготовку є неминучими, незалежно від того, який рівень кваліфікації у користувача. В таких випадках треба окремо зазначити витрати на елемент підготовки. Початкові кошти не включають кошти, витрачені на універсальне встаткування (таке як комп’ютери, DVD-програвачі і т.д.). Однак якщо в ньому є потреба, це слід зазначити. В цілому, визначте: кошти на будь-яку специфічну підготовку; кошти на інспекторські керівництва; технічні матеріали; комплект зразків або довідкових матеріалів; початкові витрати на програмне забезпечення і т.д. |  |
| 6.3.2 | **Поточні витрати**  Де доцільно, зазначте поточні витрати на проведення та підрахунок результатів окремо від витрат на тлумачення (дивіться пункти 6.4.1 – 6.5).  Цей пункт відноситься до поточних експлуатаційних витрат на використання інструменту. Він має містити кошти на матеріали інструменту (бланки для відповідей, одноразові чи багаторазові брошури з питаннями, профільні листи, коди розблокування комп’ютерних програм і т.д.) на людину і на одне проведення тесту. Майте на увазі, що в більшості випадків для проведення тестів бланкового типу такі матеріали доступні не поштучно, а постачаються пачками по 10, 25 чи 50 штук.  Складіть список ліцензійних платежів (коди розблокування програмного забезпечення, де треба) на рік або на людину, ціни на продаж або оренду багаторазових матеріалів, та кошти на одноразові матеріали на одну людину. |  |
| 6.4.1 | **Ціни на звіти, що генеруються встановленим користувальницьким ПО** |  |
| 6.4.2 | **Ціни на звіти від сервісного центру, отримувані по пошті або факсом** |  |
| 6.4.3 | **Ціни на звіти від сервісного центру, отримувані через Інтернет** |  |
| 6.5 | **Ціни на інші послуги сервісного центру: виправлення або покращення автоматичного звітування** |  |
| 6.6 | **Кваліфікація використання тесту, необхідна за твердженням поставника** *(оберіть всі підходящі варіанти)*  Цей пункт стосується наявності у користувача кваліфікації, яка вимагається поставником. Якщо поставник надав інформацію щодо кваліфікації користувача, треба зробити відповідну помітку навпроти підходящої категорії. Якщо вимоги до кваліфікації не ясні, про це слід повідомити позначивши варіант «Інше», а не «Не потрібна». «Не потрібна» значить, що мається чітке твердження з приводу відсутності необхідності в певній кваліфікації.  Для уточнення деталей щодо стандартів 2го рівню ЄФПА, перевірте останню версію даних відомостей на сайті ЄФПА. | * Не потрібна * Акредитація на проведення певного тесту * Атестат компетентності в області тестування загальних досягнень: вимірювання максимальної продуктивності (еквівалентно 2му рівню ЄФПА) * Атестат компетентності в області тестування загальних здібностей та профпридатності: вимірювання потенційної максимальної продуктивності (еквівалентно 2му рівню ЄФПА) * Атестат компетентності в області тестування загальних характеристик особистості та оцінки: вимірювання типової поведінки, установок та вподобань (еквівалентно 2му рівню ЄФПА) * Інше (впишіть): |
| 6.7 | **Професійна кваліфікація, необхідна для використання інструменту** *(помітьте всі підходящі варіанти)*  Цей пункт стосується наявності у користувача кваліфікації, яка вимагається поставником. Якщо поставник надав інформацію щодо кваліфікації користувача, треба зробити відповідну помітку навпроти підходящої категорії. Якщо вимоги до кваліфікації не ясні, про це слід повідомити позначивши варіант «Інше», а не «Не потрібна». «Не потрібна» значить, що мається чітке твердження з приводу відсутності необхідності в кваліфікації.  Для уточнення деталей щодо критеріїв користування тестів від ЄФПА, перевірте останню версію даних відомостей на сайті ЄФПА. | * Не потрібна * Практичний психолог з кваліфікацією у відповідній сфері застосування * Практичний психолог * Психолог-дослідник * Науковий дослідник не в сфері психології * Практикуючий фахівець в суміжних професіях (терапія, медицина, консультування, освіта, кадрові ресурси). Впишіть: … * Користувач тестів з кваліфікаційним рівнем ЄФПА 1 (або місцевий аналог) * Користувач тестів з кваліфікаційним рівнем ЄФПА 2 (або місцевий аналог) * Спеціаліст з тестування з кваліфікаційним рівнем ЄФПА 3 (або місцевий аналог) * Інше (впишіть): |

**ЧАСТИНА 2: ОЦІНКА ІНСТРУМЕНТУ**

**Джерела інформації**

Всього існує чотири потенційних джерел інформації, з якими можна звірятися під час проведення оцінки:

1. Керівництво та/або звіти, котрими видавець забезпечує користувача:

Вони завжди постачаються видавцем/дистриб’ютором перед тим як інструмент буде затверджено організацією-рецензентом, та є осередком матеріалу рецензії.

1. Відкрита інформація, яку можна отримати з наукової та іншої літератури:

Зазвичай ці джерела рецензент знаходить самостійно. Він може використати знайдену інформацію в рецензії, оцінивши інструмент як такий, що посилається (чи ні) на інформаційні джерела в своєму керівництві.

1. Інформація, якою володіє видавець, що ніколи офіційно не публікується або розповсюджується:

Дистриб’ютор/видавець може відкрити до неї доступ на початку, або надіслати її, коли рецензія відіслана назад до видавця на перевірку точності фактів. Рецензенту слід використати цю інформацію, однак зазначити дуже чітко на початку коментарів до технічної інформації, що «помічена зіркою оцінка в цій рецензії є посиланням на матеріали від видавця/дистриб’ютора, які [зазвичай] не постачаються користувачам тестів». Якщо в них міститься цінна інформація, в загальній оцінці слід порекомендувати видавцю публікувати ці звіти, та/або робити їх доступними для тих, хто купляє тести.

1. Конфіденційна комерційна інформація:

В деяких випадках видавці можуть мати технічно важливий матеріал, котрий вони не хочуть робити загальнодоступним з комерційних причин. На практиці існує дуже мало способів захисту інтелектуальної власності розробників тесту (хіба що авторське право). До такої інформації можуть відноситися звіти, що покривають розробку певних алгоритмів розрахунку, процедури вироблення тесту або завдання, а також технологія генерації звіту. Якщо вміст таких звітів може стати у нагоді для формування висновків у рецензії, асоціація чи організація, що несе відповідальність за рецензування, може запропонувати укласти тайну угоду з видавцем. Цей договір зв’язав би рецензентів та редактора. Рецензент тоді був би спроможний оцінити інформацію та прокоментувати технічні аспекти, у загальній оцінці відмітивши, що «помічена зіркою оцінка в цій рецензії є посиланням на матеріали від видавця/дистриб’ютора, що були розглянуті рецензентами на комерційній та конфіденційній основі. Вони не постачаються кінцевим користувачам».

**Пояснення оцінок**

Всі сегменти оцінені за допомогою наступної рейтингової системи (дивіться таблицю нижче). Детальні описи є опорними точками для кожної з оцінок.

Якщо оцінки [ 0 ] чи [ 1 ] надаються атрибутові, котрий вважається *критичним* для безпечного використання інструменту, в рецензії слід радити використовувати інструмент лише висококваліфікованим спеціалістам у виняткових обставинах або у дослідженнях.

Рецензія інструменту має вказувати на те, які його технічні якості є виключно важливими, приймаючи до уваги характер інструменту та передбачувану мету використання. Оцінки таких якостей повинні виділятися жирним шрифтом.

У нижченаведених сегментах, загальні оцінки адекватності інформації щодо валідності, надійності та норм за замовчуванням виділяються жирним шрифтом.

***Будь-який інструмент з однією чи більше оцінкою типу [ 0 ] чи [ 1 ] відносно атрибутів, котрі вважаються критичними для безпечного використання цього інструмента, буде розглядатися як такий, що не відповідає мінімальним стандартам.***

|  |  |
| --- | --- |
| **Оцінка** | **Пояснення\*** |
| [n/a] | Даний атрибут не має значення |
| 0 | Неможливо оцінити, так як інформація відсутня, або її недостатньо |
| 1 | Незадовільно |
| 2 | Задовільно |
| 3 | Добре |
| 4 | Відмінно |

\* Цю п’ятибальну шкалу було встановлено ЄФПА, проте кожний користувач може зчіплювати бали шкали, наприклад, поєднуючи бали 3 та 4 в один. Є тільки одна вимога -- між оцінками “незадовільно” (або гірше) та “задовільно” (або краще) треба провести чітке розмежування. Замість чисел можна використовувати описові вирази або символи, такі як зірки або смайлики. Якщо п’ятибальна шкала замінена чи кастомізована, користувач повинен пояснити, яким чином бали пов’язані з позначеннями п’ятибальної шкали ЄФПА.

**7. Якість пояснення теоретичної основи та поданої інформації**

В цьому сегменті треба надати ряд оцінок різноманітним аспектам або атрибутам документації, доставленої разом із інструментом. Словом «документація» покриваються всі матеріали, що надаються або вже доступні кваліфікованому користувачеві: наприклад, керівництво для інспекторів; технічні довідники; брошури з нормами; додатки до керівництва; актуальні дані від видавця/поставників, і так далі.

Поставників просять надати повний комплект таких матеріалів кожному Рецензентові. Якщо ви підозрюєте, що в надісланій вам для рецензії інформації немає якогось матеріалу, який мають надавати користувачам, зв’яжіться зі своїм редактором.

**7.1 Якість пояснення теоретичної основи**

Якщо інструмент – комп’ютерний адаптивний тест, то особливу увагу треба приділити пунктам 7.1.1 – 7.1.6.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Пункти, що оцінюються як n/a чи від 0 до 4** | | **Оцінка** | | | | | |
| 7.1.1 | Теоретичні основи конструктів | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.1.2 | Процедура розробки тесту (і/або переклад чи адаптація) | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.1.3 | Ретельність аналізу завдань та модель аналізу завдань | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.1.4 | Докази та підтвердження валідності тесту | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.1.5 | Короткий виклад актуальних досліджень | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.1.6 | **Загальний бал якості пояснення теоретичної основи**  Загальний бал визначається на підставі оцінок, наданих за пункти 7.1.1 – 7.1.5 | n/a | 0 | 1 | 2 | 3 | 4 |

**7.2 Адекватність доступної користувачеві документації (довідник користувача, технічне керівництво, додатки з нормами і т.д.)**

Тут фокус спрямований на якість охоплення необхідної інформації в документації, доступній кваліфікованим користувачам. Зверніть увагу, що підпункт 7.2 оцінює повноту і зрозумілість документації, що доступна користувачеві (довідник користувача, технічне керівництво, додатки з нормами і т.д.) -- наскільки обширно вона охоплює всю необхідну інформацію, та які пояснення надає. Якість інструмента, про яку можна судити з документації, детально розглядається в пунктах 7.1, 7.3, 9, 10 та 11.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Пункти, що оцінюються як n/a чи від 0 до 4 (надані критерії оцінки «відмінно»)** | | **Оцінка** | | | | | |
| 7.2.1 | Теоретична основа (дивіться оцінку 7.1.6)  Відмінно: Чіткий та логічний опис того, що інструмент призначений вимірювати, і чому він був сконструйований саме так. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.2.1 | Розробка  Відмінно: Детальні відомості про джерела для завдань інструменту, розробку стимульного матеріалу за прийнятими нормативами (наприклад, Haladyna, Downing, & Rodriguez, 2002; Moreno, Martinez, & Muñiz, 2006), доведення до стадії пілот-проекту, аналіз завдань, порівняльний аналіз та зміни зроблені під час випробувань на етапі розробки. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.2.2 | Розробка тесту через переклад/адаптацію  Відмінно: Інформація в керівництві демонструє, що процес перекладу/адаптації пройшов згідно з міжнародними нормами (ITC, 2000), та складався з:   * Внеску з боку носіїв нової мови * Неодноразової експертизи з боку лінгвістів та контент-експертів * Зворотного перекладу з нової мови на оригінальну * Прийняття до уваги лінгвокультурних особливостей | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.3 | Стандартизація  Відмінно: Надана чітка та детальна інформація про розміри та джерела вибірки стандартизації, а також про самий процес. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.4 | Норми  Відмінно: Надана чітка та детальна інформація про розміри та джерела груп норм, репрезентативність, умови оцінювання і т.д. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.5 | Надійність  Відмінно: Гарне обґрунтування надійності та пояснення стандартної помилки вимірювання (СПВ), а також широкий діапазон вимірювань внутрішньої узгодженості, тимчасової стійкості та підсумкові СПВ, надані з поясненням їх важливості, а також можливість генералізації оцінювального інструменту. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.6 | Конструктна валідність  Відмінно: Відмінне обґрунтування конструктної валідності з широким спектром досліджень, чітко та належним чином описаних. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.7 | Критеріальна валідність  Відмінно: Відмінне обґрунтування критеріальної валідності з широким спектром досліджень, чітко та належним чином описаних. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.8 | Згенеровані комп’ютером звіти  Відмінно: Надана чітка та детальна інформація щодо формату, мети, надійності та валідності комп’ютерних звітів. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.2.9 | **Адекватність документації, доступної користувачеві (довідник користувача, технічне керівництво, додатки з нормами і т.д.)**  Цей бал визначається на підставі оцінок, наданих за пункти 7.2.1 – 7.2.8 | n/a | 0 | 1 | 2 | 3 | 4 |

**7.3 Якість методичних інструкцій для користувача**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Пункти, що оцінюються як n/a чи від 0 до 4 (надані критерії оцінки «відмінно»)** | | **Оцінка** | | | | | |
| 7.3.1 | Щодо проведення тестування  Відмінно: Надана чітка та детальна інформація, а також покроковий процедурний довідник, з докладними вказівками щодо того, як відповідати на питання учасників та вирішувати будь-які виникаючі під час процесу проблеми. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.2 | Щодо підрахунку результатів тесту  Відмінно: Надана чітка та детальна інформація з описами перевірочних випробувань, в яких повідомляється як впоратись із можливими помилками під час підрахунку.  Якщо підрахунок результатів здійснюється за допомогою комп’ютера, чи наявні докази того, що підрахунок зроблено вірно? | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.3 | Щодо нормування  Відмінно: Надана чітка та детальна інформація з описами перевірочних випробувань, в яких повідомляється як впоратись із можливими помилками в нормуванні.  Якщо процедура нормування відбувається на комп’ютері, чи наявні докази того, що перетворення шкальних балів правильне, і що задіяна потрібна група норм? | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.4 | Щодо тлумачення та звітування  Відмінно: Вичерпні поради щодо тлумачення різних показників, розуміння нормативних вимірювань, та врахування співвідношення різних шкал з наочним прикладом та розбором ситуацій; також з порадами з приводу того, як обходитись з можливим впливом протиріч у відповідях, стилі відповідей, шахрайством, і т.д. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.5 | Щодо зворотного зв’язку та проведення опитування протестованих та інших  Відмінно: Вичерпні поради щодо того, як надати якісний зворотний зв’язок учасникам тесту, включаючи варіант з використанням знегерованих комп’ютером звітів (якщо вони доступні). | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.6 | Щодо інформації з приводу проблем, пов’язаних зі справедливістю та упередженістю  Відмінно: Детальна інформація про дослідження гендерної та етнічної упередженості, з відповідними застереженнями щодо використання та узагальнення валідностей. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.7 | Обмеження в використанні  Відмінно: Чітка інформація з приводу того, хто може, а хто не може оцінюватися, з вичерпним обґрунтуванням вирішень щодо обмежень (наприклад, через різні типи інвалідності, чи через вимоги певного рівню грамотності і т.д.) | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.8 | Програмне забезпечення та технічна підтримка  Відмінно: У випадку комп’ютерного чи інтернет-тестування: повна інформація щодо вимог до браузеру, встановлення будь якого необхідного програмного забезпечення, та його експлуатації (включаючи також можливі помилки та покриваючи різні системи), а також щодо доступності технічної підтримки. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.9 | Посилання та допоміжні матеріали  Відмінно: Детальні посилання на важливу допоміжну наукову літературу та інші матеріали про оцінювальні інструменти. | n/a | 0 | 1 | 2 | 3 | 4 |
| 7.3.10 | **Якість методичних інструкцій для користувача**  Цей бал визначається на підставі оцінок, наданих за пункти 7.3.1 – 7.3.9 | n/a | 0 | 1 | 2 | 3 | 4 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 7.4 | **Загальна адекватність**  Загальний бал за сегмент 7 визначається на підставі оцінок, наданих за підсекції 7.1, 7.2 та 7.3. | n/a | 0 | 1 | 2 | 3 | 4 |

|  |
| --- |
| **Коментарі рецензентів щодо документації:** (коментарі щодо теоретичних основ та поданої інформації) |
|  |

**8. Якість тестових матеріалів**

**8.1 Якість матеріалів тесту бланкового типу**

*(можна пропустити, якщо тест відноситься до іншого типу)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Пункти, що оцінюються як n/a чи від 0 до 4** | | **Оцінка** | | | | | |
| 8.1.1 | Загальна якість тестових матеріалів (тестові брошури, бланки для відповідей, тестовий об’єкт) | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.1.2 | Наскільки легко учасник тесту може зрозуміти завдання | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.1.3 | Ясність та повнота інструкцій (включаючи зразки завдань) для учасника тесту | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.1.4 | Наскільки легко учасник тесту може відповісти на питання чи виповнити завдання | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.1.5 | Наскільки якісно сформульовані завдання/питання, та наскільки зрозумілий графічний вміст у тому випадку, якщо тест невербальний | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.1.6 | **Якість матеріалів тесту бланкового типу**  Цей бал визначається на підставі оцінок, наданих за пункти 8.1.1 – 8.1.5 | n/a | 0 | 1 | 2 | 3 | 4 |

**8.2 Якість матеріалів комп’ютерних та Інтернет-тестів**

*(можна пропустити, якщо тест відноситься до іншого типу)*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Пункти, що оцінюються як n/a чи від 0 до 4** | | **Оцінка** | | | | | |
| 8.2.1 | Якість дизайну програмного забезпечення (наприклад, стійкість до помилок в тому випадку, коли натискаються неправильні клавіші, або до збою підключення до інтернету, і т.д.) | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.2.2 | Наскільки легко учасник тесту може зрозуміти завдання | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.2.3 | Ясність та повнота інструкцій (включаючи зразки завдань) для учасника тесту; наприклад, щодо того як працювати в програмі та давати відповіді, якщо тест проводиться на комп’ютері | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.2.4 | Наскільки легко учасник тесту може відповісти на питання чи виповнити завдання | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.2.5 | Якість дизайну користувальницького інтерфейсу | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.2.6 | Захист тесту від неавторизованого доступу до завдань чи відповідей | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.2.7 | Наскільки якісно сформульовані завдання/питання, та наскільки зрозумілий графічний вміст у тому випадку, якщо тест невербальний | n/a | 0 | 1 | 2 | 3 | 4 |
| 8.2.8 | **Якість матеріалів комп’ютерних та Інтернет-тестів**  Цей бал визначається на підставі оцінок, наданих за пункти 8.2.1 – 8.2.7 | n/a | 0 | 1 | 2 | 3 | 4 |

|  |
| --- |
| **Коментарі рецензентів щодо якості матеріалів** |
|  |

**9. Норми**

**Загальне керівництво з присвоєння оцінок у цьому розділі**

Досить складно встановити чіткі критерії оцінювання технічних якостей інструменту. Цей матеріал надає вказівки щодо того, які величини можуть оцінюватися як незадовільні, задовільні, хороші та відмінні. Однак, цей матеріал призначений лише для ознайомлення. Сутність інструменту, сфера його застосування, якість даних, на яких засновуються норми, та види рішень, для яких він використовуватиметься – це все впливає на те, як присвоюються оцінки.

Для осмислення первинних показників тесту можна виділити два способи шкалювання або категоризації первинних показників (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). По-перше, набір шкальованих оцінок або норм може отримуватися з первинних показників референтної групи. Він називається нормативно-орієнтованою інтерпретацією (див. пункт 9.1). По-друге, стандарти можуть бути отримані зі сфери здібностей чи предмету, яким потрібно оволодіти (змістовна інтерпретація), або прохідні бали можуть бути отримані з результатів емпіричного дослідження валідності (критеріально-орієнтована інтерпретація) (див. пункт 9.2). У двох останніх випадках первинні дані поділяються на дві (наприклад, "залік" або "незалік") або більше категорії різних показників, наприклад, для того, щоби назначати пацієнтам різні програми лікування в залежності від показників, чи назначити учням, що отримали показники нижче критичного, корекційне навчання, або приймати чи відмовляти кандидатам при відборі персоналу.

**9.1 Нормативно-орієнтована інтерпретація**

*(цю підсекцію можна пропустити, якщо даний вид інтерпретації не застосовується)*

**Примітки щодо міжнародних норм**

Придатність міжнародних норм (одномовних) потребує ретельного розгляду. Якщо вони були ретельно встановлені з вибірок, отриманих з ряду країн, то вони повинні оцінюватися нарівні з групами норм з однієї країни (одномовними). Якщо використовується нелокальна норма, мають бути надані переконливі докази еквівалентності як тестових версій, так і вибірок, для підтвердження доцільності її використання. Загалом такі докази потребують досліджень, які демонструють скалярну відповідність між оригінальною версією та її перекладом. *Якщо вони відсутні, про це треба зазначити в «Коментарях рецензентів» у кінці 9-го сегменту.*

Міжнародна норма може бути найбільш придатною для міжнародного використання (порівнюючи людей, що пройшли тест на різних мовах), але потрібно враховувати викладені нижче проблеми при визначенні її доцільності. Як мінімум, використання міжнародної норми потребує демонстрації еквівалентності вимірювання між версіями тесту на оригінальній мові та в перекладі.

Характер вибірки

* Баланс джерел вибірки (наприклад, вибірка, 95% якої – німці, 2% -- італійці, та 3% -- британці, не може вважатися дійсним міжнародним стандартом). Спеціальні процедури зважування вибірки можуть бути проведені задля кращого відображення її різних складових.
* Еквівалентність вихідних даних (кваліфікація, освіта, умови тестування тощо) різних частин вибірки. Нормативні вибірки, що не дозволяють це оцінити, непридатні.

Тип вимірювання:

* Якщо тест не містить вербального змісту (або його мало), вплив на переклад відповідно буде менший. Це стосується тестів фізичних реакцій та деякою мірою тестів абстрактного та схематичного мислення, де вплив на показники має бути меншим.

Еквівалентність версії тесту, що використовується з різномовними вибірками.

* Має бути підтвердження того, що всі мовні версії коректно перекладено/адаптовано.
* Чи наявні докази того, що яка-небудь з груп склала тест не оригінальною мовою?

Схожість показників в різних вибірках:

* Мають бути наявні дані про відносні показникові патерни вибірок з різних країн. Якщо є значні розбіжності, про це має бути зазначено, а висновки щодо використання обговорені. Наприклад, якщо іспанська вибірка набирає більше балів за фактор, ніж голландська вибірка, чи надається пояснення з приводу того, як порівнювати членів кожної з груп або третьої групи з середнім показником? Чи наявне тлумачення розбіжностей?

*Потрібно прокоментувати відсутність джерел доказів у відповідній секції («Коментарі рецензентів») в кінці сегменту.*

Вказівки щодо поширення норм поза тими групами, що включені в міжнародні норми, мають бути включені до керівництва до інструменту.

* Наприклад, якщо 20% норми складають німці, 20% -- італійці, 20% -- британці, і ще 20% -- голландці, то доцільним може бути її використання в якості групи порівняння з швейцарцями та бельгійцями, але не з китайцями.

|  |  |  |
| --- | --- | --- |
| **9.1** | **Нормативно-орієнтована інтерпретація**  Якщо інструмент призначений для використання без звернення до норм чи референтних груп (наприклад, іпсативні тести призначені лише для внутрішньоіндивідуальних порівнянь), слід помічати категорію “не застосовується” замість “інформація відсутня”. Однак, рецензент має оцінити, чи виправдані причини для відсутності норм – якщо ні, треба помітити категорію “інформація відсутня”. | |
| 9.1.1 | Придатність для локального застосування – як локальних, так і міжнародних норм  Зверніть увагу, що у випадку з адаптованими тестами лише локальні (в рамках країни) чи дійсні міжнародні норми придатні для присвоєння балів від 2 до 4, навіть якщо конструкт є еквівалентним поміж різних культур. У випадку, коли трапляються проблеми інваріантності оцінювання, мають бути надані окремі норми для (під)груп, а сутність самих проблем має пояснюватися. | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня | 0 |
| Непридатні для локального застосування (наприклад, непридатні іноземні вибірки) | 1 |
| Локальні (в рамках країни) вибірки, що не дуже підходять до потрібної сфери застосування, але можуть використовуватися з обачністю | 2 |
| Локальні (в рамках країни) вибірки чи релевантні міжнародні вибірки, що достатньо придатні для передбачуваної сфери застосування | 3 |
| Локальні (в рамках країни) вибірки чи релевантні міжнародні вибірки, отримані з чітко визначених соціальних груп та придатні для передбачуваної сфери застосування | 4 |
| 9.1.2 | Придатність для передбачуваного застосування | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня | 0 |
| Норма чи норми незадовільні для передбачуваного застосування | 1 |
| Задовільні норми генеральної сукупності і (або) набір таблиць норм; норми задовільні для деяких, але не всіх сфер застосування | 2 |
| Хороший набір таблиць норм | 3 |
| Відмінний набір вибірок, вікові та статеві норми з інформацією щодо відмінностей між групами (наприклад, різними етнічними групами) | 4 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 9.1.3 | Розміри вибірки (класичне нормування)  Для багатьох цілей вибірки, які складаються з менше ніж 200 учасників тесту, будуть занадто малими, оскільки ступінь деталізації у хвостах розподілу буде надто низькою. SEсереднє для z-балу, там де N = 200, дорівнює 0.071 від SD - або краще ніж один T-бал. Хоча ступінь неточності може мати тільки незначні наслідки в центрі розподілу, вплив на хвости розподілу може бути досить значним (і це може бути діапазоном показників, що найбільш важливі для прийняття рішень на основі показників тесту). Якщо наявні міжнародні норми, то загалом, через їх гетерогенність, вони повинні бути більшими, ніж це передбачено типовими вимогами до локальних вибірок.  Для низьких та високих ставок передбачаються різні кількісні дані. Загалом, застосування високих ставок – це коли значне рішення основане принаймні частково на показниках тесту. | | | |
|  | Низькі ставки | Високі ставки |  |
| Даний атрибут не застосовується до інструменту | | | n/a |
| Інформація відсутня | | | 0 |
| Розмір вибірки незадовільний | напр., <200 | напр., 200-299 | 1 |
| Розмір вибірки задовільний | напр., 200-299 | напр., 300-399 | 2 |
| Хороший розмір вибірки | напр., 300-999 | напр., 400-999 | 3 |
| Відмінний розмір вибірки | напр., ≥ 1000 | напр., ≥ 1000 | 4 |
| 9.1.4 | Розміри вибірки безперервного нормування  Безперервне нормування останнім часом набирає все більшої популярності. Зокрема, воно використовується для тестів, які призначені для шкіл (наприклад, групи від 1 до 8 початкової школи) або для конкретної вікової групи (наприклад, тест розумових здібностей для дітей віком 6-16 років). Безперервне нормування більш ефективне, тому що менше число респондентів потрібне для отримання такої ж точності норм. Бехгер, Хемкер та Маріс (Bechger, Hemker, & Maris 2009) розрахували певні значення розміру безперервних груп норм, що забезпечать однакову точність у порівнянні з класичним нормуванням. Коли використовуються вісім підгруп *N* = 70 (8x70) дає таку ж точність як і *Ns* 200 (8x200) за класичним підходом; *N* = 100 (x8) порівнюється з 300 (x8) та *N* = 150 (x8) з 400 (x8). У цих випадках точність на основі підходу безперервного нормування навіть краща в середніх групах, але дещо гірша у зовнішніх групах. Ще однією перевагою окрім більшої ефективності є те, що можна вирахувати значення, основані на регресивній лінії, для проміжних груп норм. Однак, підхід засновується на достатньо точному статистичному припущенні. Автор тесту має показати, що ці припущення підтвердилися, або що відхилення від цих припущень не впливатиме на точність норм.  Майте на увазі, що коли кількість груп більша, кількість респондентів в кожній групі може бути менша або навпаки. Для важливих рішень, таких як зарахування до школи, необхідна кількість зміщується на один в сторону збільшення. | | | |
| Даний атрибут не застосовується до інструменту | | | n/a |
| Інформація відсутня | | | 0 |
| Розмір вибірки незадовільний (наприклад, менше восьми підгруп з максимальною кількістю в 69 респондентів в кожній з них) | | | 1 |
| Розмір вибірки задовільний (наприклад, вісім підгруп, в кожній з яких по 70-99 респондентів) | | | 2 |
| Хороший розмір вибірки (наприклад, вісім підгруп, в кожній з яких по 100-149 респондентів) | | | 3 |
| Відмінний розмір вибірки (наприклад, вісім підгруп, в кожній з яких якнайменш 150 респондентів) | | | 4 |
| 9.1.5 | Процедури, що використовуються при обранні вибірки *(помітьте один варіант)*  Група норм має бути репрезентативною для групи, що розглядається. Вибірка може вважатися репрезентативною щодо передбачуваної сукупності, якщо склад вибірки по відношенню до перемінних (наприклад, вік, стать, освіта) схожий зі складом сукупності, *і* коли вибірка збирається за допомогою моделі імовірнісної вибірки. В такій моделі шанс бути включеним у вибірку однаковий для кожного елемента сукупності. Можна використовувати різні методи як для імовірнісної, так і неімовірнісної вибірки.  У випадку з імовірнісною вибіркою, коли особа є одиницею вибору, можна виділити три методи: повністю випадковий, систематичний (наприклад, кожний десятий член сукупності) або розшарований (для деяких важливих перемінних, наприклад таких, як стать, кількість, яку потрібно відібрати, є фіксованою, аби гарантувати репрезентативність цих перемінних). Однак (наприклад, заради ефективності), вибірки можуть бути сформовані з груп осіб (наприклад, шкільні класи) або ж можна застосовувати комбінацію групового та індивідуального вибіркового методу. У випадку з неімовірнісною вибіркою виділяють чотири методи: вибірка виключної зручності (просто додайте кожну протестовану особу до групи норми, як це робиться в більшості вибірок для відбору кандидатів на працевлаштування; дані, отримані за результатами, можуть класифікуватися в змістовні підгрупи, що основані на біографічній та ситуативній інформації), вибірка по квотам (те ж що і у вибірці виключної зручності, але заздалегідь зазначається скільки потрібно респондентів у кожній підгрупі, як це робиться у дослідженнях громадської думки), вибірка за методом сніжного кому (ви просите свої друзів прийняти участь у тестуванні, потім просите їх попрохати своїх друзів, і так далі) та цільова вибірка (наприклад, для участі відбираються протилежні групи). | | | |
| Інформація відсутня | | | [ ] |
| Імовірнісна вибірка – випадкова | | | [ ] |
| Імовірнісна вибірка – систематична | | | [ ] |
| Імовірнісна вибірка – розшарована | | | [ ] |
| Імовірнісна вибірка – кластерна | | | [ ] |
| Імовірнісна вибірка – багатофазова (напр., перший кластер, а потім випадковий з-поміж інших кластерів) | | | [ ] |
| Неімовірнісна вибірка – зручності | | | [ ] |
| Неімовірнісна вибірка – по квотам | | | [ ] |
| Неімовірнісна вибірка – «сніжний ком» | | | [ ] |
| Неімовірнісна вибірка – цільова | | | [ ] |
| Інші, опишіть: … | | | [ ] |
| 9.1.6 | Репрезентативність вибірки норм | | | |
| Даний атрибут не застосовується до інструменту | | | n/a |
| Інформація відсутня | | | 0 |
| Репрезентативність недостатня для передбаченої сфери застосування, або репрезентативність не може бути об’єктивно оцінена з наданої інформації | | | 1 |
| Задовільно | | | 2 |
| Добре | | | 3 |
| Відмінно: Дані зібрані за допомогою моделі випадкового вибору; наданий всебічний опис відбору вибірки (вибірок) та сукупностей в плані основних перемінних (таких як стать, вік, освіта, культурне походження, сфера зайнятості); встановлена гарна репрезентативність даних перемінних | | | 4 |
| 9.1.7 | Якість наданої інформації щодо відмінностей між групами меншин, впливу віку, статі, і т.д. | | | |
| Даний атрибут не застосовується до інструменту | | | n/a |
| Інформація відсутня | | | 0 |
| Інформація незадовільна | | | 1 |
| Загальна інформація задовільна, присутній мінімальний аналіз | | | 2 |
| Хороші описи та аналіз груп та відмінностей між ними | | | 3 |
| Відмінний аналіз та всебічний розгляд нюансів тлумачення | | | 4 |
| 9.1.8 | Як давно проводилися нормативні дослідження? | | | |
| Даний атрибут не застосовується до інструменту | | | n/a |
| Інформація відсутня | | | 0 |
| Незадовільно – 20 років назад чи більше | | | 1 |
| Задовільно – нормам від 15 до 19 років | | | 2 |
| Добре – нормам від 10 до 14 років | | | 3 |
| Відмінно – нормам менш ніж 10 років | | | 4 |
| 9.1.9 | Результати практики (істотно лише для тесту фізичних реакцій) | | | |
| Даний атрибут не застосовується до інструменту | | | n/a |
| Інформація відсутня, проте можна очікувати певних результатів | | | [ ] |
| Надана загальна інформація | | | [ ] |
| Норми для наступного застосування тесту після проходження інтервалу повторного тестування | | | [ ] |

**9.2 Критеріально-орієнтована інтерпретація**

*(цю підсекцію можна пропустити, якщо даний вид інтерпретації не застосовується)*

Для того, щоби визначити критичну оцінку, можна провести відмінність між такими процедурами, коли використовуються судження експертів (такі методи також називають змістовним нормуванням, див. пункт 9.2.1) та процедурами, де використовуються фактичні дані відносно зв’язку між тестовим показником та зовнішнім критерієм (критеріально-орієнтоване нормування, див. пункт 9.2.2).

|  |  |  |
| --- | --- | --- |
| **9.2.1** | **Змістовне нормування** | |
| 9.2.1.1 | Якщо для визначення критичної оцінки використовуються експертні судження, то чи належним чином відібрані та навчені судді?  Судді повинні бути обізнаними щодо інформаційної області тесту та методу визначення норми, а також належним чином навченими для оцінки роботи учасника тесту. Мають бути описані методика відбору суддів та рекомендації щодо навчання. | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня або її недостатньо | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 9.2.1.2 | Якщо для визначення критичної оцінки використовуються експертні судження, то чи достатня кількість суддів?  Необхідна кількість суддів залежить від завдань та контексту. Передбачувана кількість має розглядатися як абсолютний мінімум. | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня або її недостатньо | 0 |
| Незадовільно (менше двох) | 1 |
| Задовільно (двоє суддів) | 2 |
| Добре (три судді) | 3 |
| Відмінно (чотири чи більше суддів) | 4 |
| 9.2.1.3 | Якщо для визначення критичної оцінки використовуються експертні судження, то який метод визначення норми вказується? *(один варіант)* | |
| Метод Недільского | [ ] |
| Метод Ангоффу | [ ] |
| Метод Ебеля | [ ] |
| Метод Лівінгстона (метод порогових груп) | [ ] |
| Метод Берка (метод контрастних груп) | [ ] |
| Метод Беука | [ ] |
| Метод Гофстея | [ ] |
| Інший, опишіть: | [ ] |
| 9.2.1.4 | Якщо для визначення критичної оцінки використовуються експертні судження, який метод визначення надійності взаємного тестування вказується? | |
| Коефіцієнт ро | [ ] |
| Коефіцієнт Каппа | [ ] |
| Коефіцієнт Лівінгстон | [ ] |
| Коефіцієнт Бреннан та Кейн | [ ] |
| Внутрішньокласовий коефіцієнт кореляції | [ ] |
| Інший, опишіть: | [ ] |
| 9.2.1.5 | Якщо для визначення критичної оцінки використовуються експертні судження, яким є розмір коефіцієнту надійності взаємного тестування (наприклад, Каппа чи внутрішньокласового коефіцієнту кореляції)?  В науковій літературі немає чітких однозначних стандартів інтерпретації цих типів коефіцієнтів, але значення нижче .60, як правило, вважаються недостатніми. Нижче використовується класифікація Шраута (Shrout 1998). Однак, дана класифікація має використовуватися дещо обачливо, адже долі та апріорні дані можуть впливати на значення коефіцієнту Каппа. | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня | 0 |
| Незадовільно (наприклад, *r* < 0.60) | 1 |
| Задовільно (наприклад, 0.60 ≤ *r* < 0.70) | 2 |
| Добре (наприклад, 0.70 ≤ *r* < 0.80) | 3 |
| Відмінно (наприклад, *r* ≥ 0.80) | 4 |
| 9.1.2.6 | Як давно проводилися нормативні дослідження? | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня | 0 |
| Незадовільно – 20 років назад чи більше | 1 |
| Задовільно – нормам від 15 до 19 років | 2 |
| Добре – нормам від 10 до 14 років | 3 |
| Відмінно – нормам менш ніж 10 років | 4 |
| 9.2.1.7 | Результати практики (істотно лише для тесту фізичних реакцій) | |
| Інформація відсутня, проте можна очікувати певних результатів | [ ] |
| Надана загальна інформація | [ ] |
| Норми для наступного застосування тесту після звичайного інтервалу повторного тестування | [ ] |
| **9.2.2** | **Критеріально-орієнтоване нормування** | |
| 9.2.2.1 | Якщо критична оцінка базується на емпіричному дослідженні, якими є результати та якість даного дослідження?  Для відповіді на це запитання не надається певних чітких вказівок з приводу допустимого рівня співвідношення, не тільки через те, що для кожного прогнозованого критерію стандарти можуть варіюватися, але також тому, що на результати прогнозування будуть впливати інші змінні, такі як апріорні дані чи долі. Тому рецензент має покладатися на власні професійні знання та компетентність для надання оцінки. Також треба взяти до уваги склад вибірки що використовувалась для дослідження(чи подібний він до групи, для якої призначений тест; більш гетерогенний, чи більш гомогенний?) та розмір цієї групи. | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 9.2.2.2 | Як давно проводилися нормативні дослідження? | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня | 0 |
| Незадовільно – 20 років назад чи більше | 1 |
| Задовільно – нормам від 15 до 19 років | 2 |
| Добре – нормам від 10 до 14 років | 3 |
| Відмінно – нормам менш ніж 10 років | 4 |
| 9.2.2.3 | Результати практики (істотно лише для тесту фізичних реакцій) | |
| Інформація відсутня, проте можна очікувати певних результатів | [ ] |
| Надана загальна інформація | [ ] |
| Норми для наступного застосування тесту після проходження інтервалу повторного тестування | [ ] |
| 9.3 | **Загальна відповідність вимогам**  Загальний бал визначається на підставі оцінок, наданих за пункти 9.1 – 9.2.2.3.  Загальна оцінка нормативно-орієнтованої інтерпретації не може бути вищою за оцінку, надану в пункті про розмір вибірки, але може бути нижчою через іншу доступну інформацію. Особливо важливою є інформація з приводу репрезентативності та актуальності норм. Якщо використовується детермінована вибірка, якість норм може бути оцінена максимум як «задовільна», але тільки в тому випадку, якщо опис груп норм показує, що розподіл на важливі перемінні подібний досліджуваній сукупності. Загальна оцінка має відображати характеристики найзначніших та найбільших норм, аніж «середнє» поміж всіх опублікованих норм.  Загальна оцінка критеріально-орієнтованої інтерпретації не може бути вищою за оцінку розміру коефіцієнту взаємного тестування, але може бути нижчою через іншу доступну інформацію. Особливо важливою є інформація з приводу належного застосування методу та його якості, рекомендацій щодо спеціального навчання, та кількості суддів. Якщо критична оцінка грунтується на емпіричному дослідженні, оцінка не може бути вищою за ту, що надається в пункті 9.2.2.1, але вона може бути нижчою в тому випадку, коли дослідження надто старі. | |
| Даний атрибут не застосовується до інструменту | n/a |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |

|  |
| --- |
| **Коментарі рецензентів з приводу норм:** Невеликий звіт щодо норм та їх походження, що включає в себе інформацію про заходи, яких було вжито автором або видавцем для оновлення норм на постійній основі. Коментарі стосовно нелокальних норм також мають бути тут. |
|  |

**10. Надійність**

**Загальне керівництво з присвоєння оцінок для цього розділу**

Надійність співвідноситься із тим, наскільки результати вільні від дисперсії похибки вимірювання (тобто діапазону очікуваної похибки вимірювання). Ці вказівки основані на необхідності мати невелику середньоквадратичну помилку для оцінки надійності. Орієнтовні критерії надійності надані у відношенні до двох окремих контекстів: використання інструментів для прийняття рішень щодо груп людей (наприклад, організаційний діагноз) та для проведення індивідуального оцінювання. Вимоги до надійності останнього суворіші, ніж до першого. Інші фактори також можуть впливати на вимоги щодо надійності, такі як вид рішень, що приймаються та те, чи інтерпретуються шкали самі по собі або ж поєднуються з іншими в композитну шкалу. В останньому випадку при оцінці фокус повинен бути на надійності цього поєднання, а не на надійності компонентів.

Якщо інструмент було перекладено та/або адаптовано з нелокального контексту, можна застосувати докази надійності оригінальної версії для підтвердження якості перекладеної/адаптованої версії. В цьому випадку потрібно надати докази еквівалентності між вимірюванням на новій мові та вимірюванням на оригінальній мові. Без цього неможливо узагальнити дані в одній країні/мовній версії для іншої. Однак для внутрішньої узгодженості перевага надається доказам щодо надійності, які базуються на локальних групах, адже такі докази точніші та їх зазвичай легше отримати. Переглянути вказівки щодо встановлення еквівалентності можна у вступі до сегменту про Валідність. Пам’ятку ключових пунктів для коментування у випадку, коли інструмент було перекладено та/або адаптовано з нелокального контексту, включено до Додатку.

Досить складно встановити чіткі критерії оцінювання технічних якостей інструменту. Цей матеріал надає вказівки щодо того, які величини можуть оцінюватися як незадовільні, задовільні, добрі чи відмінні. Однак цей матеріал призначений лише для ознайомлення. Сутність інструменту, сфера його застосування, якість даних, на яких засновуються норми, та види рішень, для яких він використовуватиметься – це все впливає на те, як присвоюються оцінки. За певних умов надійність з коефіцієнтом 0.70 прийнятна, за інших -- ні. Саме тому підсумкові оцінки повинні грунтуватися на Ваших, як рецензента, судженнях, висновках та експертних знаннях, а не вилучатися за допомогою усереднення ряду оцінок.

Щоби надати уявлення про діапазон та розподілення величин, що пов'язані з різними шкалами, які утворюють інструмент, зазначте *кількість шкал* в кожній секції. Наприклад, якби інструмент, що використовується для рішень групового рівня, мав 15 шкал, з яких п'ять мали б коефіцієнт надійності повторного тестування нижче, ніж 0.6, шість з них -- між 0.60 та 0.70, а інші чотири в діапазоні від 0.70 до 0.80, середнє значення стійкості оцінювалося б як “задовільне” (так як середнє 15 величин потрапляє саме в цю категорію). Якщо мова йде більш ніж про одне дослідження, спочатку вираховується середнє значення величини на шкалу з урахуванням розміру вибірки; в деяких випадках можуть бути доступні результати з мета-аналізу, вони мають оцінюватися так само. Це записується наступним чином:

|  |  |  |
| --- | --- | --- |
| **Стійкість** | Кількість шкал  *(якщо доцільно)* | M\* |
| Інформація відсутня | [ - ] | 0 |
| Задовільно (наприклад,*r* < 0.60) | [ 5 ] | 1 |
| Незадовільно (наприклад, 0.60 ≤ *r* < 0.70) | [ 6 ] | 2 |
| Добре (наприклад, 0.70 ≤ *r* < 0.80) | [ 4 ] | 3 |
| Відмінно (наприклад, *r* ≥ 0.80) | [ 0 ] | 4 |

Величини для кожного можливого прикладу оцінок указуються виключно з ознайомчою метою -- особливо розмежування між "задовільно", "добре" та "відмінно". Для рішень, що мають велике значення, таких як відбір кандидатів на працевлаштування, ці приклади величин будуть на .10 більше. Однак слід пам’ятати, що рішення часто базуються на сукупних шкальованих показниках. Сукупності можуть мати значно більшу надійність, ніж їх складові первинні показники. Наприклад, первинні показники в багатошкальному інструменті можуть мати коефіцієнт надійності приблизно в 0.70, в той час як коефіцієнт надійності другорядних сукупностей показників Великої п’ятірки, основаних на них, буде близько 0.90. В хороших посібниках до тестів буде зазначено як про вторинні, так і про первинні показники.

Зрозуміло, що в багатьох випадках неможливо вирахувати фактичне середнє значення. Тут вимагається Ваша найбільш точна оцінка з урахуванням інформації, яку можна знайти в документах. Є місце для коментарів. Можете ділитися будь-якими зауваженнями щодо точності Ваших оцінок. Наприклад, у деяких випадках про надто високий рівень внутрішньої узгодженості може бути зауважено, що він вказує на «роздутий концепт».

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 10 | **Надійність** | | | | |
| 10.1 | Інформація щодо надійності *(можна обрати два варіанти)* | | | | |
| Інформація відсутня | | | | [ ] |
| Надано лише один коефіцієнт надійності (для кожного фактору чи субфактору) | | | | [ ] |
| Надано лише одну оцінку стандартної помилки вимірювання (для кожного фактору чи субфактору) | | | | [ ] |
| Коефіцієнти надійності для ряду різних груп (для кожного фактору чи субфактору) | | | | [ ] |
| Стандартна помилка вимірювання для ряду різних груп (для кожного фактору чи субфактору) | | | | [ ] |
| 10.2 | **Внутрішня узгодженість**  Недоцільно використовувати коефіцієнти внутрішньої узгодженості для оцінки надійності тестів на швидкість, гетерогенних шкал (у тому числі згадуються емпіричні та критеріальні шкали; Cronbach, 1970), показників впливу на навколишнє середовище (Nunnally & Bernstein, 1994), та емерджентних властивостей (Schneider & Hough, 1995). В таких випадках всі атрибути, що відносяться до внутрішньої узгодженості, мають бути позначені як такі, що “не застосовуються”. Цей метод також не підходить для визначення надійності іпсативних шкал. Для таких типів шкал більше підійде оцінка альтернативних форм чи повторне тестування.  Коефіцієнти внутрішньої узгодженості оцінюють надійність краще, ніж метод розщеплення, відкоригований формулою Спірмана-Брауна. Тому застосування методу розщеплення виправдано лише в тому випадку, коли інформація з приводу відповідей на окремі запитання недоступна з будь-яких причин. Про метод розщеплення можна зазначити в пункті 10.7 (Інші методи). | | | | |
| 10.2.1 | Розмір вибірки | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | 0 |
| Одне дослідження, що не відповідає вимогам (напр., розмір вибірки менше 100) | | | | 1 |
| Одне дослідження, що відповідає вимогам (напр., розмір вибірки 100-200) | | | | 2 |
| Одне велике дослідження (напр., розмір вибірки більший за 200), або більше одного дослідження, що відповідає вимогам | | | | 3 |
| Хороший набір великих досліджень і досліджень, що відповідають вимогам | | | | 4 |
| 10.2.2 | Зазначений тип коефіцієнтів *(помітьте всі підходящі)* | | | | |
| Не застосовується | | | | n/a |
| Коефіцієнт альфа або KR-20 | | | | [ ] |
| Лямбда-2 | | | | [ ] |
| Найбільша нижня грань | | | | [ ] |
| Омега (факторний аналіз) | | | | [ ] |
| Тета (факторний аналіз) | | | | [ ] |
| Інший, опишіть: … | | | | [ ] |
| 10.2.3 | Розмір коефіцієнтів | | | Кількість факторів  *(якщо доцільно)* | M\* |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | [ ] | 0 |
| Незадовільно (напр., r < 0.70) | | | [ ] | 1 |
| Задовільно (напр., 0.70 ≤ r < 0.80) | | | [ ] | 2 |
| Добре (напр., 0.80 ≤ r < 0.90) | | | [ ] | 3 |
| Відмінно (напр., r ≥ 0.90) | | | [ ] | 4 |
| 10.2.4 | Коефіцієнти надійності з вибірками, котрі… *(один варіант)* | | | | |
| …не відповідають передбачуваним учасникам тесту, та призводять до придатніших коефіцієнтів (напр., збільшення дисперсії штучною гетерогенністю) | | | | [ ] |
| …не відповідають передбачуваним учасникам тесту, але вплив розміру коефіцієнтів не ясний | | | | [ ] |
| …не відповідають передбачуваним учасникам тесту, та призводять до менш придатних коефіцієнтів (напр., зменшення дисперсії обмеженням діапазону) | | | | [ ] |
| ...відповідають передбачуваним учасникам тесту | | | | [ ] |
| Не застосовується | | | | n/a |
| **10.3** | **Надійність повторного тестування – часова стійкість**  Повторне тестування відноситься до відносно коротких часових інтервалів, тоді як часова стійкість відноситься до довших інтервалів, під час яких більша кількість змін є прийнятною. Обидва аспекти особливо важливі для тестів, що використовуються для прогнозування на довші періоди. Для оцінки часової стійкості може потребуватися більш ніж одне повторне тестування.  Проведення повторного тестування недоцільне для оцінки надійності вимірювань стану (взагалі, високий коефіцієнт повторного тестування зробив би характер тесту стану недійсним). В такому випадку, на всі пункти, що стосуються надійності повторного тестування, треба відповідати “*не застосовується*”. | | | | |
| 10.3.1 | Розмір вибірки | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | 0 |
| Одне дослідження, що не відповідає вимогам (напр., розмір вибірки менше 100) | | | | 1 |
| Одне дослідження, що відповідає вимогам (напр., розмір вибірки 100-200) | | | | 2 |
| Одне велике дослідження (напр., розмір вибірки більший за 200), або більше одного дослідження, що відповідає вимогам | | | | 3 |
| Хороший набір великих досліджень і досліджень, що відповідають вимогам | | | | 4 |
| 10.3.2 | Розмір коефіцієнтів | | | Кількість факторів  *(якщо доцільно)* | M\* |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | [ ] | 0 |
| Незадовільно (напр., r < 0.60) | | | [ ] | 1 |
| Задовільно (напр., 0.60 ≤ r < 0.70) | | | [ ] | 2 |
| Добре (напр., 0.70 ≤ r < 0.80) | | | [ ] | 3 |
| Відмінно (напр., r ≥ 0.80) | | | [ ] | 4 |
| 10.3.3 | Дані щодо інтервалу повторного тестування *(оберіть підходящий варіант чи впишіть часовий інтервал повторного тестування)* | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | [ ] |
| Інтервал: ………… | | | | ... |
| 10.3.4 | Коефіцієнти надійності з вибірками, котрі… *(один варіант)* | | | | |
| …не відповідають передбачуваним учасникам тесту, та призводять до придатніших коефіцієнтів (напр., збільшення дисперсії штучною гетерогенністю) | | | | [ ] |
| …не відповідають передбачуваним учасникам тесту, але вплив розміру коефіцієнтів не ясний | | | | [ ] |
| …не відповідають передбачуваним учасникам тесту, та призводять до менш придатних коефіцієнтів (напр., зменшення дисперсії обмеженням діапазону) | | | | [ ] |
| ...відповідають передбачуваним учасникам тесту | | | | [ ] |
| Не застосовується | | | | n/a |
| 10.4 | Надійність еквівалентності (паралельних чи альтернативних форм) | | | | |
| 10.4.1 | Розмір вибірки | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | 0 |
| Одне дослідження, що не відповідає вимогам (напр., розмір вибірки менше 100) | | | | 1 |
| Одне дослідження, що відповідає вимогам (напр., розмір вибірки 100-200) | | | | 2 |
| Одне велике дослідження (напр., розмір вибірки більший за 200), або більше одного дослідження, що відповідає вимогам | | | | 3 |
| Хороший набір великих досліджень і досліджень, що відповідають вимогам | | | | 4 |
| 10.4.2 | Чи підтверджуються припущення щодо паралелізму\* різних версій тесту, надійність еквівалентності котрих досліджується?  \*Майте на увазі, що тести можуть вважатися паралельними, якщо в одній групі середні бали, відхилення та кореляція з іншими тестами одні й ті самі. | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | 0 |
| Задовільно | | | | 1 |
| Незадовільно | | | | 2 |
| Добре | | | | 3 |
| Відмінно | | | | 4 |
| 10.4.3 | Розмір коефіцієнтів | Кількість факторів  *(якщо доцільно)* | | | M\* |
| Не застосовується | | | | n/a |
| Інформація відсутня | [ ] | | | 0 |
| Задовільно (напр., r < 0.70) | [ ] | | | 1 |
| Незадовільно (напр., 0.70 ≤ r < 0.80) | [ ] | | | 2 |
| Добре (напр., 0.80 ≤ r < 0.90) | [ ] | | | 3 |
| Відмінно (напр., r ≥ 0.90) | [ ] | | | 4 |
| 10.4.4 | Коефіцієнти надійності з вибірками, котрі… *(один варіант)* | | | | |
| …не відповідають передбачуваним учасникам тесту, та призводять до придатніших коефіцієнтів (напр., збільшення дисперсії штучною гетерогенністю) | | | | [ ] |
| …не відповідають передбачуваним учасникам тесту, але вплив розміру коефіцієнтів не ясний | | | | [ ] |
| …не відповідають передбачуваним учасникам тесту, та призводять до менш придатних коефіцієнтів (напр., зменшення дисперсії обмеженням діапазону) | | | | [ ] |
| ...відповідають передбачуваним учасникам тесту | | | | [ ] |
| Не застосовується | | | | n/a |
| **10.5** | **Метод на основі теорії відповідей на завдання (Item Response Theory, далі -- IRT)** | | | | |
| 10.5.1 | Розмір вибірки  Якщо надійність перевіряється методами на основі IRT, досить складно встановити універсальні норми розміру вибірок, оскільки вимоги до них будуть різними в залежності від того, який формат чи модель відповідей на завдання використовуються. В залежності від моделі відповідей мінімальні значення для “задовільних” розмірів вибірок наступні: 200 для однопараметричних досліджень, 400 для двопараметричнх досліджень та 700 для трьохпараметричних (за Parshall, Davey, Spray, & Kalohn, 2001). Дані величини застосовуються до дихотомічних моделей, але рецензент може спиратися на них при оцінюванні політомних моделей, розмір вибірок яких може бути меншим. | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | 0 |
| Одне дослідження, що не відповідає вимогам | | | | 1 |
| Одне дослідження, що відповідає вимогам | | | | 2 |
| Одне велике дослідження, або більше одного дослідження, що відповідає вимогам | | | | 3 |
| Хороший набір великих досліджень і досліджень, що відповідають вимогам | | | | 4 |
| 10.5.2 | Тип зазначених коефіцієнтів *(помітьте всі підходящі варіанти)*  Перший метод визначає надійність досліджуваної латентної властивості, що в IRT замінює істинний бал, себто тестовий бал (див. Embertson & Reise, 2000). Другий метод базується на інформації про окремі завдання та надає оцінку надійності в разі задоволення типових для IRT вимог (Mokken, 1971). Третій метод надає оцінку точності вимірювання рівня латентної властивості. | | | | |
| Достовірність досліджуваної латентної властивості | | | | [ ] |
| Ро | | | | [ ] |
| Інформаційна функція | | | | [ ] |
| Інше, опишіть: | | | | [ ] |
| Не застосовується | | | | n/a |
| 10.5.3 | Розмір коефіцієнтів (основаних на довжині фінального тесту)  Обидва нормативи з коефіцієнтів надійності (включаючи ро) надані по відношенню до інформаційної функції. Нормативи інформаційної функції основані на нормативних величинах коефіцієнтів надійності, оскільки Інформація = 1/SE2, та, враховуючи деякі досить часті припущення, *r* = 1 - SE2. Майте на увазі, що SE та інформаційні показники залежать від показника латентної властивості, а також те, що кожен тест має діапазон, в якому інформаційний показник є оптимальним. Оцінка не має бути основана на цьому оптимальному показнику апріорі, а має -- на інформаційному показнику балу або ряду балів особливої важливості (наприклад, критичні бали). Для таких балів інформаційний показник може бути оптимальним, але необов'язково. Якщо таких балів немає, тоді оцінка основується на середній величині інформаційного показника (див. також Reise & Havilund, 2005р.). Рекомендується використовувати ці практичні правила з обережністю через відсутність належного досвіду їх використання. | | | Кількість факторів  *(якщо доцільно)* | M\* |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | [ ] | 0 |
| Незадовільно (напр., r < 0.70; інформація < 3.33) | | | [ ] | 1 |
| Задовільно (напр., 0.70 ≤ r < 0.80; 3.33 ≤ інформація < 5.00) | | | [ ] | 2 |
| Добре (напр., 0.80 ≤ r < 0.90; 5.00 ≤ інформація < 10.00) | | | [ ] | 3 |
| Відмінно (напр., r ≥ 0.90; інформація ≥ 10.00) | | | [ ] | 4 |
| **10.6** | **Надійність взаємного оцінювання**  Якщо підрахунок результатів тесту не включає в себе процесу винесення суджень (наприклад, просто підсумовуючи бали тестів множинного вибору), цей тип надійності не вимагається, і всі пункти, що відносяться до надійності взаємного оцінювання, мають бути позначені як такі, що “не застосовуються”. Майте на увазі, що хоч надійність взаємного оцінювання може не застосовуватися до цілого тесту загалом, вона може застосовуватися до одного чи більше субтестів (наприклад, деякі субтестів тесту інтелектуальних здібностей). | | | | |
| 10.6.1 | Розмір вибірки | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | 0 |
| Одне дослідження, що не відповідає вимогам (напр., розмір вибірки менше 100) | | | | 1 |
| Одне дослідження, що відповідає вимогам (напр., розмір вибірки 100-200) | | | | 2 |
| Одне велике дослідження (напр., розмір вибірки більший за 200), або більше одного дослідження, що відповідає вимогам | | | | 3 |
| Хороший набір великих досліджень і досліджень, що відповідають вимогам | | | | 4 |
| 10.6.2 | Тип зазначених коефіцієнтів (помітьте всі підходящі варіанти) | | | | |
| Не застосовується | | | | n/a |
| Процентна узгодженість | | | | [ ] |
| Коефіцієнт Каппа | | | | [ ] |
| Внутрішньокласовий коефіцієнт кореляції | | | | [ ] |
| Коефіцієнт Йота | | | | [ ] |
| Інший, опишіть: | | | | [ ] |
| 10.6.3 | Розмір коефіцієнтів  До деяких методів, згаданих в пункті 10.6.2, зазначені числа можуть не підходити, якщо не вираховується *r*. | Кількість факторів  *(якщо доцільно)* | | | M\* |
| Не застосовується | | | | n/a |
| Інформація відсутня | [ ] | | | 0 |
| Незадовільно (напр., r < 0.60) | [ ] | | | 1 |
| Задовільно (напр., 0.60 ≤ r < 0.70) | [ ] | | | 2 |
| Добре (напр., 0.70 ≤ r < 0.80) | [ ] | | | 3 |
| Відмінно (напр., r ≥ 0.80) | [ ] | | | 4 |
| **10.7** | **Інші методи оцінки надійності** | | | | |
| 10.7.1 | Розмір вибірки | | | | |
| Не застосовується | | | | n/a |
| Інформація відсутня | | | | 0 |
| Одне дослідження, що не відповідає вимогам (напр., розмір вибірки менше 100) | | | | 1 |
| Одне дослідження, що відповідає вимогам (напр., розмір вибірки 100-200) | | | | 2 |
| Одне велике дослідження (напр., розмір вибірки більший за 200), або більше одного дослідження, що відповідає вимогам | | | | 3 |
| Хороший набір великих досліджень і досліджень, що відповідають вимогам | | | | 4 |
| 10.7.2 | Опишіть метод: | | | | |
| 10.7.3 | Результати | | Кількість факторів  *(якщо доцільно)* | | M\* |
| Не застосовується | | | | n/a |
| Інформація відсутня | | [ ] | | 0 |
| Незадовільно | | [ ] | | 1 |
| Задовільно | | [ ] | | 2 |
| Добре | | [ ] | | 3 |
| Відмінно | | [ ] | | 4 |
| **10.8** | **Загальна відповідність вимогам**  Загальний бал визначається на підставі оцінок, наданих за пункти 10.1 – 10.7.3. *Не усереднюйте оцінки для отримання загального балу.*  Внутрішня узгодженість може бути недоречною для деяких інструментів (базові риси чи агрегати шкал), і в таких випадках слід зосередитись на інформації щодо повторного тестування. В інших випадках (оцінювання стану), надійність повторного тестування не буде доречною, так що слід звернути увагу на внутрішню узгодженість. Для фінального оцінювання також прийміть до уваги:  -- для чого використовується тест: для індивідуального оцінювання чи для прийняття рішень, що стосуються групи людей  -- сутність рішення (низькі чи високі ставки)  -- скільки типів дослідження надійності зазначено (один чи більше)  -- чи наявні стандартні помилки вимірювання  -- процедурні питання, наприклад, розмір групи, кількість досліджень надійності, гетерогенність груп(и), коефіцієнт якої вираховується, кількість експертів (якщо визначається надійність взаємного тестування), довжина інтервалу повторного тестування, і т.д.  -- всеосяжність звітів досліджень на визначення надійності. | | | | |
| Інформація відсутня | | | | 0 |
| Незадовільно | | | | 1 |
| Задовільно | | | | 2 |
| Добре | | | | 3 |
| Відмінно | | | | 4 |

|  |
| --- |
| **Коментарі рецензентів щодо надійності:** Підкресліть сильні та слабкі аспекти наявних доказів надійності. Коментарі стосовно генералізації інформації щодо надійності та еквівалентності також мають бути тут (якщо доцільно). |
|  |

**11. Валідність**

**Загальне керівництво з присвоєння оцінок для цього розділу**

Валідність -- це ступінь відповідності результатів поставленим завданням та первинній меті тесту. В спеціальній літературі розрізняють велику кількість типів валідності, наприклад, Дрент та Сійтсма (Drenth & Sijtsma, 2006, p. 334-340) згадують вісім різних типів. Відмінності можуть бути пов'язані з метою валідності або з процесом валідизації за допомогою спеціальних методик аналізу даних. За останні десятиліття минулого століття все більше посилювалась думка про те, що валідність потрібно розглядати як єдине поняття, а також те, що диференціацію в типах валідності потрібно розглядати виключно як різні способи збору доказів (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Борсбум, Мелленберг та ван Херден (Borsboom, Mellenbergh, & Van Heerden, 2004) стверджують, що тест є валідним для вимірювання атрибуту, якщо варіація атрибуту дає варіацію у вимірюваних результатах.

Незалежно від того, якому типові валідності надається перевага, для стандартизованого судження потрібно трохи систематизувати поняття валідності. Для цього маємо дві окремих підсекції -- для конструктної та критеріальної валідностей. В залежності від мети тесту, один з цих аспектів валідності може бути доречнішим за інший. Але при цьому зрозуміло, що конструктна валідність є більш фундаментальним поняттям, і докази критеріальної валідності можуть сприяти при встановленні конструктної валідності тесту. Також зрозуміло, що тест може мати різні валідності в залежності від виду рішень, які приймаються за допомогою тесту, виду використаних вибірок тощо. Однак, невід'ємним принципом рецензії тесту є те, що одне судження складається про якість (конструктної або критеріальної) валідності тесту. Це судження повинно відображати якість доказу, що підтримує твердження про те, що тест може бути використаним для інтерпретацій, вказаних у довіднику. Чим ширша сфера застосування тесту, тим більше доказів валідності має надати автор/видавець. Зверніть увагу, що остаточні оцінки конструктної та критеріальної валідності будуть чимось на кшталт “середнього значення” цих доказів, а також що можуть бути ситуації або групи, для яких тест може мати більшу або меншу валідність (або для яких валідність взагалі може не досліджуватися).

Якщо інструмент було перекладено та/або адаптовано з нелокального контексту, потрібно надати докази еквівалентності вимірювання новою мовою відносно оригіналу. Без цього неможливо узагальнити дані в одній країні/мовній версії щодо іншої. Приклади доказів еквівалентності:

* Інваріантність структури конструкту – наприклад, у вигляді структури фактору або кореляції зі стандартними показниками.
* Схожа критеріальна валідність – наприклад, схожий профіль кореляцій багатофакторного інструменту з незалежним зовнішнім критерієм, таким як оцінки відповідності займаній посаді.
* Завдання демонструють схожі шаблони наповнення шкал, наприклад, завдання співвідносяться з іншими шкалами за одним шаблоном; найсильніше/найслабше наповнення завданнями схоже як в оригінальному, так і в перекладеному варіанті.
* Учасники тесту, що володіють двома мовами, мають однакові профілі в обох мовах (надійність альтернативної форми).

Генералізація валідності потребує більш переконливих доказів при здійсненні перекладу тестів між різними мовними сім’ями (наприклад, з індоєвропейських мов на мови семітської групи). В такій ситуації еквівалентність знаходиться під значною загрозою через культурні відмінності та розбіжності в структурі мови. Однак, генералізація валідності може бути отримана з доказів інваріантності валідності попередніх перекладів, якщо тест було перекладено на декілька мов. Наприклад, якщо тест на шведській мові було перекладено на французьку, німецьку та італійську мови, і еквівалентність цих варіантів тесту було доказано.

Розглядаючи питання еквівалентності, доцільно притримуватися класифікації Ван де Вівера та Пуртінга (Van de Vivjer & Poorting 2005):

* Структурна / функціональна еквівалентність

Наявний доказ того, що оригінальна та перекладена версії вимірюють одні й ті самі психологічні конструкти в усіх групах. Це загалом ілюструється однаковістю паттернів кореляції між змінними в групах.

* Еквівалентність одиниць вимірювання

Наявний доказ того, що одиниці вимірювання однакові, але джерела серед груп різні (тобто індивідуальні розбіжності, виявлені в групі А, можна порівняти з розбіжностями, що виявлені в групі Б, але абсолютні первинні показники для груп А і Б не можна прямо порівняти, попередньо не провівши перешкалювання).

* Скалярні / повні показники еквівалентності

Одна й та сама одиниця вимірювання та одне й те саме джерело (тобто первинні бали мають однакове значення та можуть бути порівняні між групами).

Критерії для оцінки та нотатки в підсекціях 11.1 та 11.2 надають деякі орієнтири щодо того, які величини можна оцінювати як незадовільні, задовільні, хороші та відмінні. Однак, цей матеріал призначений лише для ознайомлення. Сутність інструменту, сфера його застосування, якість даних, на яких засновуються оцінки валідності, та види рішень, для яких він використовуватиметься - це все впливає на те, як присвоюються оцінки. Критерії розміру вибірок базуються на аналізі потужності розмірів вибірки, що потрібен для виявлення валідностей помірної величини в разі їх наявності.

**11.1 Конструктна валідність**

Метою конструктої валідності є дати відповідь на запитання про те, чи дійсно тест вимірює передбачений конструкт або ж, частково чи в більшій мірі, щось інше. Існують наступні загальноприйняті методи дослідження конструктної валідності: оцінка факторної структури чи конфірматорний факторний аналіз, кореляційний аналіз тестових завдань, порівняння середніх показників груп, в яких можуть очікуватися розбіжності, тестування на інваріантність факторної структури та диференціальне функціонування завдань (DIF -- differential item functioning) для різних груп, кореляції з іншими інструментами, які призначені для тестування того ж конструкта (конвергентна валідність) або іншого (дискримінанта валідність), підхід “множинні ознаки -- множинні методи” (MTMM -- Multitrait-multimethod matrix), методологія теорії відповідей на тестові завдання (IRT) та (квазі-)експериментальний підхід.

|  |  |  |
| --- | --- | --- |
| 11.1 | Конструктна валідність | |
| 11.1.1 | Яка методика використовується (помітьте всі підходящі варіанти) | |
| Інформація відсутня | [ ] |
| Оцінка факторної структури | [ ] |
| Конфірматорний факторний аналіз | [ ] |
| (Скориговані) кореляції тестових завдань | [ ] |
| Тестування інваріантності структури та диференціальне функціонування завдань поміж груп | [ ] |
| Відмінності між групами | [ ] |
| Кореляція з іншими інструментами та критерій ефективності функціонування | [ ] |
| Кореляція MTMM | [ ] |
| Методологія теорії відповідей на тестові завдання (IRT) | [ ] |
| (Квазі-)експериментальний підхід | [ ] |
| Інше, опишіть: | [ ] |
| 11.1.2 | Чи обгрунтовують результати факторного аналізу (конфірматорного чи оцінки факторної структури) структуру тесту? | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.3 | Чи достатньо питання/завдання корелюють з (суб)тестовою оцінкою?  Зверніть увагу, що дуже високі кореляції можуть вказувати на те, що питання більш-менш тотожні, і що вимірюваний концепт може бути дуже вузьким (так званий “роздутий концепт”). | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.4 | Чи є факторна структура незмінною в усіх групах та чи не містить тест упередження в питаннях/твердженнях (DIF)?  Це дослідження може бути проведено на основі моделей класичної теорії тесту або загальних принципах теорії відповідей на тестові завдання (IRT). Якщо виявлено упередження в питаннях чи завданнях, необхідно оцінити його вплив на загальну суму балів (невеликий вплив є прийнятним). | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.5 | Чи відмінності поміж середніх балів цільових груп такі, як очікувалось?  Наприклад, очікується, що за тест арифметичних здібностей учні групи №8 отримають вищий бал ніж учні групи №6; діти з діагнозом СДУГ отримають вищий результат за тест на гіперактивність ніж діти, в яких немає цього діагнозу; торгівці отримають вищий результат за тест на комерційні знання ніж інше робоче населення. Хоч результати і відповідають очікуваному, такий тип досліджень не дозволяє зробити чітких висновків щодо конструктної валідності тесту. Тим не менш, сутність такого дослідження полягає в тому, що якщо очікувана різниця не виявляється, це має викликати серйозні сумніви у конструктній валідності тесту. | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.6 | Середнє значення та діапазон кореляцій між тестом та тестами, які вимірюють подібні конструкти  Кореляція балів тесту зі шкалами схожих інструментів, так звана конгруентність або конгруентна валідність, є ключовим елементом процесу валідизації конструкту. Нормативні значення коефіцієнтів конгруентної валідності слід трактувати гнучко. У випадку, коли було корельовано два дуже схожі інструменти (водночас із отриманими даними), коефіцієнт кореляції, який можна оцінити як “задовільний” дорівнював би 0.60 чи більше. У випадках, коли інструменти менш схожі або коли сеанси проведення тесту розділені певним проміжком часу, нижчі показники можуть вважатися задовільними. При оцінці конгруентної валідності високі кореляції слід тлумачити з обачністю. Коли кореляції перевищують 0.90, існує імовірність того, що шкали вимірюють один і той самий конструкт. Якщо дані шкали утворюють нову шкалу та сталу мітку, то проблеми немає. Але це було би проблемою в тому випадку, якщо ці шкали повинні були додати корисну варіацію до того, що вже вимірюють інші шкали. Надані нормативи торкаються кореляцій, що не пристосовані для методу варіантів або регресійного поділу. Тому, оцінюючи коефіцієнти конгруентної валідності, також необхідно враховувати надійність обох інструментів. Наприклад, якщо коефіцієнт надійності обох інструментів -- .75, то максимальна кореляція між інструментами дорівнюватиме .56. Якщо показники надійності вищі, то кореляції також будуть вищі. | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.7 | Чи демонструє кореляція з іншими інструментами хорошу дискримінантну валідність в плані конструктів, які не вимірюються тестом? | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.8 | Якщо використовується підхід “множинних ознак -- множинних методів”, то чи підтверджують результати конструктну валідність тесту (чи дійсно він вимірює те, що має вимірювати, а не щось інше)?  Зверніть увагу: при використанні підходу MTMM може відпасти необхідність в дослідженнях, що згадуються в пунктах 11.1.6 та 11.1.7. | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.9 | Інше, наприклад методологія IRT чи (квазі-)експериментальний підхід (опишіть): | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.10 | Розмір вибірок  Вказані нижче норми стосуються досліджень на принципах класичної теорії тесту. Для оцінки параметрів питань/завдань в рамках методології IRT “задовільні” розміри вибірок такі: більш ніж 200 для однопараметричних досліджень, більш ніж 400 для двопараметричних досліджень та більш ніж 700 для трьохпараметричних досліджень (за Parshall, Davey, Spray, & Kalohn, 2001). | |
| Інформація відсутня | 0 |
| Незадовільно (напр., розмір вибірки менший ніж 100) | 1 |
| Задовільно (напр., розмір вибірки -- 100-200) | 2 |
| Добре (напр | 3 |
| Відмінно | 4 |
| 11.1.11 | Якість інструментів як критеріїв чи маркерів | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 11.1.12 | Як давно проводилися дослідження валідності?  Складно сформулювати загальне правило для прийняття віку дослідження до уваги. Для тестів, що вимірюють конструкти у тій сфері, в якій відбулись деякі важливі теоретичні розробки, дослідження, що проводились 15 років тому, можуть бути майже зовсім непридатними, тоді як для інших тестів дослідження, що проводились 20 років назад (і давніші) є доречними і зараз. | |
| Кількість років | ……. |
|
|
|
|
| 11.1.13 | **Конструктна валідність - загальна відповідність вимогам**  Загальний бал визначається на підставі оцінок, наданих за пункти 11.1.1 -- 11.1.12. *Не усереднюйте оцінки для отримання загального балу.*  Окрім результатів дослідження конструктної валідності, для заключного рішення Ви також маєте прийняти до уваги чи правильно використовуються техніки аналізу (наприклад, чи була застосована поправка значень рівня значущості для кореляції інструменту з іншими інструментами без чітких гіпотез), чи подібні вибірки дослідження до груп, для яких призначається тест (наприклад, більша гетерогенність значно підвищить кореляції, вибірки студентів можуть дати результати, котрі не можна буде узагальнити), а також розмір досліджених вибірок, якість інших інструментів, що використовувались (наприклад, в дослідженні конвергентної та дискримінантної валідності) та давність проведених досліджень. | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |

**11.2 Критеріальна валідність**

Доказ критеріальної валідності (поточної та прогностичної валідності) співвідноситься з дослідженнями, де практичні критеріальні міри були скорельовані зі шкалами. Прогностичні дослідження зазвичай відносяться до ситуацій, в яких оцінювання проводилось у “якісно відмінний” від часу критеріального вимірювання момент (наприклад, у контексті праці, для відбіркового оцінювання призначеного спрогнозувати успішність трудової діяльності інструмент використали би під час відбору), а не просто є питанням того, наскільки довгим був часовий інтервал між інструментом та критеріальним дослідженням. Дослідження також можуть бути “постдиктивними”, наприклад, якщо результати потенційного відбіркового тесту корелюються з раніше отриманими претендентом на посаду оцінками, наданими його безпосереднім керівництвом. Загалом, докази критеріальної валідності необхідні для будь-яких типів тестів. Тим не менш, якщо в керівництві чітко зазначено, що тест не використовується задля прогностичних цілей (як, наприклад, навчальні тести що оцінюють успішність), то критеріальна валідність може не оцінюватися взагалі.

|  |  |  |
| --- | --- | --- |
| 11.2 | Критеріальна валідність | |
| 11.2.1 | Тип критеріального дослідження чи досліджень (помітьте всі підходящі варіанти) | |
| Прогностичний | [ ] |
| Поточний | [ ] |
| Постдиктивний | [ ] |
| 11.2.2 | Розмір вибірок | |
| Інформація відсутня | 0 |
| Одне дослідження, що не відповідає вимогам (напр., розмір вибірки менше 100) | 1 |
| Одне дослідження, що відповідає вимогам (напр., розмір вибірки 100-200) | 2 |
| Одне велике дослідження (напр., розмір вибірки більший за 200), або більше одного дослідження, що відповідає вимогам | 3 |
| Хороший набір великих досліджень і досліджень, що відповідають вимогам | 4 |
| 11.2.3 | Якість критеріальних вимірювань | |
| Інформація відсутня | 0 |
| Якість задовільна | 1 |
| Якість незадовільна | 2 |
| Хороша якість | 3 |
| Відмінна якість в плані надійності та відображення критеріального конструкту | 4 |
| 11.2.4 | Сила зв’язку між тестом та критеріями  Досить складно встановити чіткі критерії оцінювання розміру коефіцієнтів критеріальної валідності інструменту. Критеріальна валідність з коефіцієнтом 0.20 може бути вельми доцільною в окремих ситуаціях, в той час як коефіцієнт 0.40 -- малоцінним в інших. Коефіцієнт 0.30 може вважатися достатнім для відбору персоналу, тоді як у контексті навчання коефіцієнти зазвичай вищі. З цих причин оцінки мають грунтуватися на Ваших, як рецензента, судженнях, висновках та експертних знаннях, а не вилучатися за допомогою усереднення кореляційних коефіцієнтів. Дані нормативи засновані на роботі Хемфіла (Hemphill, 2003; також див. Meyer et al., 2001) та стосуються кореляцій, яких не було скориговано для регресійного поділу як у предикторі, так і в критерії. Однак коефіцієнти можуть бути скориговані для звуження діапазону.  Надані нижче діапазони стосуються коефіцієнтів валідності, тому що кореляції між тестами та критеріями є найбільш популярним способом відображення критеріальної валідності. Однак в клінічних ситуаціях дані щодо чутливості та специфічності критерію можуть сказати більше про зв’язок між тестом та критерієм. ROC-криві є популярним методом визначення чутливості та специфічності. Светс (Swets, 1998) надає короткий огляд значень ROC-кривих в різних сферах. Для деяких типів медичного діагнозу значення знаходяться між .81 та .97, для випробувань на детекторі брехні -- .70-.95, для оцінки навчальних досягнень (залік/не залік) -- .71-.94. Дані величини можна використовувати в якості орієнтиру, але рецензент має вирішити сам, покладаючись на свої знання та досвід, чи | |
| Інформація відсутня | 0 |
| Незадовільно (*r* < 0.20) | 1 |
| Задовільно (0.20 ≤ *r* < 0.35) | 2 |
| Добре (0.35 ≤ *r* < 0.50) | 3 |
| Відмінно (*r* ≥ 0.50) | 4 |
| 11.2.5 | Як давно проходили дослідження валідності?  Складно сформулювати загальне правило для прийняття віку дослідження до уваги. Для тестів що призначені прогнозувати поведінку у стрімко мінливому світі, дослідження, що проводились 15 років тому, можуть бути майже зовсім непридатними, тоді як для інших тестів дослідження, що проводились 20 років назад (і давніші) є доречними і зараз. | |
| Кількість років | ... |
| 11.2.6 | **Критерільна валідність -- Загальна відповідність вимогам**  Загальний бал визначається на підставі оцінок, наданих за пункти 11.2.1 -- 11.2.5. *Не усереднюйте оцінки для отримання загального балу.*  Окрім результатів дослідження критеріальної валідності, для заключного рішення Ви також маєте прийняти до уваги чи правильно використовуються процедури та техніки аналізу (наприклад, чи присутня контамінація критерію, корекція для регресійного поділу, перехресна перевірка), чи подібні вибірки дослідження до груп, для яких призначається тест (наприклад, поправка на звуження діапазону), а також розмір досліджених вибірок, якість інших критеріальних інструментів, що використовувались (наприклад, чи не присутня нестача критерію) та давність проведених досліджень. | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |

**11.3 Загальна валідність**

При оцінці загальної валідності слід пам’ятати про значущість конструктної валідності як найкращого індикатора того, чи вимірює тест саме те, що він має вимірювати. В деяких випадках головним свідченням цього можуть бути критеріальні дослідження. Такий інструмент може мати оцінку «задовільно» чи вище за критеріальну валідність, та нижчу за «задовільно» за конструктну валідність. Загалом, оцінка за Загальну валідність буде ідентичною оцінці чи за Конструктну валідність, чи за Критеріальну валідність, в залежності від того, яка вище. Однак, в залежності від мети тесту один з цих типів свідчення може вважатися доцільнішим за інший. До оцінки за Загальну валідність не треба підходити як до середнього показнику або найменшого загального знаменнику.

|  |  |  |
| --- | --- | --- |
| 11.3 | **Валідність – Загальна відповідність вимогам**  Загальний бал визначається на підставі оцінок, наданих за пункти 11.1.1 -- 11.2.6. *Не усереднюйте оцінки для отримання загального балу.* | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |

|  |
| --- |
| **Коментарі рецензентів щодо валідності** (включно з доказами валідності). Коментарі стосовно генералізації валідності та еквівалентності також мають бути тут (якщо доцільно). |
|  |

**12 Якість комп’ютерних звітів**

Той факт, що багато хто з поставників, ясна річ, побажає захистити свою інтелектуальну власність у вигляді алгоритмів та правил підрахунку, ускладнює оцінку комп’ютерних звітів. На практиці необхідна інформація має бути доступною для використання в цілях рецензування у технічних довідниках, де описаний процес розробки звіту та його принцип; також її можна знайти переглядаючи зразки контрольних тестових прикладів конфігурацій підрахунку. В ідеалі документація також має містити опис процедур, що використовувались для того, щоби перевірити генерацію звіту на точність, зв’язність та релевантність. Для рецензування мають бути надані якнайменше три звіти з різними профілями розрахунку, включаючи фактичні бали, навіть якщо алгоритми генерації звітів конфіденційні.

Кожний пункт містить питання, що мають допомогти при оцінці кожного з нижченаведених атрибутів; також надане тлумачення оцінки «відмінно» (4).

|  |  |  |
| --- | --- | --- |
| **Пункти, що оцінюються як n/a чи від 0 до 4 (надані критерії оцінки «відмінно»)** | | |
| 12.1 | **Широта охоплення**  Як повнота, так і конкретика варіюються від звіту к звіту. Звіти також варіюються в плані кількості людей, яким вони підходять. В деяких випадках виробляється декілька типів звіту --для різних груп одержувачів*.*   * *Чи покриває звіт увесь ряд факторів, вимірюваних інструментом?* * *Чи робить він це на такому рівні конкретики, що може бути виправданий рівнем детальності інформації, котру можна отримати з результатів інструменту?* * *Чи виправдана глибина деталізації звіту (тобто кількість чітких груп оцінок на шкалі, що використовуються для зв’язувати різні блоки тексту в звіті) в плані похибок вимірювання шкал?* * *Чи для того ж контингенту, для якого був розроблений інструмент, зроблений звіт? (наприклад, групи для котрих придатні групи норм або критеріальні дані тощо)* | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно: Відмінне співпадіння широти охоплення інструменту та звіту, з таким рівнем конкретики в звіті, котрий цілком відповідає ступеню деталізації шкал. Розумне використання всіх результатів, що можна отримати з інструменту. | 4 |
| 12.2 | **Надійність**   * *Наскільки послідовними та несуперечливими є звіти в плані тлумачення подібних наборів оціночних даних?* * *Чи зробили зміст звіту варіантним (напр., за допомогою випадкового відбору з еквівалентних блоків тексту), та чи належним чином це було зроблено?* * *Чи виправдано тлумачення балів та різниця між ними в плані похибок шкального вимірювання?* | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно: Відмінна узгодженість та послідовність в тлумаченні та належні застереження щодо інтерпретації та рекомендацій з огляду на похибки вимірювання, що лежать в їх основі. | 4 |
| 12.3 | **Придатність або валідність**  Зв’язок між інструментом та змістом звіту може пояснюватися як у звіті, так і в окремому документі. Якщо звіт ґрунтується на клінічній оцінці, документ має містити опис процесу розробки змісту експертами, а також правила, за якими результати співвідносяться зі змістом.   * *Наскільки міцно пов’язані між собою зміст звіту та оціночні дані інструменту? До якої міри звіт відхиляється чи розходиться з інформацією, отриманою з результатів інструменту?* * *Чи є наявне чітке співвідношення змісту звіту з характеристиками, вимірюваними інструментом?* * *Чи надає він обґрунтоване пояснення щодо критеріїв, до яких такі характеристики можуть мати відношення?* * *Який емпіричний доказ надається для підтвердження наявності цього взаємозв’язку?*   Важливо розглянути як конструктну валідність звіту (тобто до якої міри він надає тлумачення що співпадає з визначенням лежачих в основі конструктів), так і критеріальну валідність (у випадку якщо зроблені твердження які можна пов’язати з емпіричними даними). | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно: Співвідношення шкал зі змістом звіту з чітким обгрунтуванням. | 4 |
| 12.4 | **Чесність, або відсутність систематичної упередженості**   * *Чи здаються зміст звіту та використана в ньому мова недоречними для певних груп людей?* * *Чи надає звіт чітке пояснення з приводу можливої упередженості в результатах інструмента?* * *Чи доступні альтернативні іншомовні версії? Якщо так, то чи належним чином була перевірена їх рівноцінність?* | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно: Чіткі зауваження та пояснення можливої упередженості, доступні в усіх необхідних мовах користувача. | 4 |
| 12.5 | **Придатність**  Ця оцінка буде в значній мірі залежати від складності мови, що використовується в звіті, складності описаних конструктів та мети, задля котрої був створений інструмент.   * *Чи підійде форма та зміст звіту одержувачам, для яких він призначається?* * *Чи відповідає мова звіту рівню грамотності та математичних здібностей передбачуваних читачів?* | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно: Дуже висока придатність, звіт цілком підходить передбачуваним одержувачам. | 4 |
| 12.6 | **Довжина**  Довжина є аспектом Практичності, тому її слід розглядати саме з цього боку при оцінці. Занадто довгі звіти можуть вказувати на надмірну інтерпретацію результатів. Саме тому довжина звітів також оцінюється. В середньому, якщо тлумачення однієї шкали сягає більше однієї сторінки, то звіт можна вважати занадто довгим. | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |
| 12.7 | **Загальна відповідність вимогам комп’ютерних звітів**  Загальний бал визначається на підставі оцінок, наданих за пункти 12.1 -- 12.6. *Не усереднюйте оцінки для отримання загального балу.* | |
| Інформація відсутня | 0 |
| Незадовільно | 1 |
| Задовільно | 2 |
| Добре | 3 |
| Відмінно | 4 |

|  |
| --- |
| **Коментарі рецензентів щодо комп’ютерних звітів**  При оцінці можна розглянути такі фактори як прийняття звітом до уваги будь-яких перевірок узгодженості та послідовності відповідей, вимірювання спотворення відповідей (наприклад, вимірювання основної тенденції в відповідях типу ранжування) та інших показників, зважаючи на які можна оцінити, наскільки впевнено можна тлумачити результати окремої протестованої людини.  Сюди також можна включити коментарі щодо складності алгоритмів, наприклад, чи розглядаються одночасно декілька шкал, як слід поводитися зі шкальними профілями, і т.д. Така складність, звичайно, має бути обґрунтована у технічному довіднику. |
|  |

**13. Остаточне оцінювання**

|  |
| --- |
| **Оціночний звіт про тест**  Цей сегмент має містити стислу аргументовану оцінку тесту. Тут треба описати його плюси та мінуси, а також надати загальні рекомендації щодо того як і коли його використовувати, разом із зауваженнями з приводу того, коли його не слід використовувати (де доцільно).  Також тут треба надати короткий виклад позитивних та негативних аспектів, пов’язаних з адаптованими чи перекладеними тестами. Перелік важливих факторів для розгляду знаходиться в Додатку як нагадування про нотатки у важливих сегментах. Коментуйте це тільки в тому випадку, коли є потреба.  Оцінка має покривати такі теми як придатність інструменту для різноманітних функцій оцінювання та сфер застосування; потреба в спеціальному навчанні чи наявності специфічних навичок; простота використання; якість та кількість інформації наданої поставником; чи є важлива інформація, що не надається користувачеві, а також чи виникають будь-які проблеми з адаптацією та перекладом інструмента (див. Додаток).  Додайте коментарі щодо досліджень, котрі здійснюються в даний час, а також плани поставника про майбутню доробку та поліпшення якості. |
|  |
| **Висновки** |
|  |

|  |  |  |
| --- | --- | --- |
| **Рекомендації** *(оберіть один варіант)*  Слід виділити найбільш підходящий пункт рекомендацій з даного списку. В нормі треба додати будь-які коментарі, обгрунтування чи застереження. Також слід додати невеликі вказівки з приводу того, в яких ситуаціях та яким чином інструмент може використовуватися, а також застереження про сфери, в яких можливе помилкове нецільове використання інструменту.  **Для того, щоби оцінити інструмент як такий, що можна рекомендувати для загального використання (4й чи 5й пункт зі списку), слід переконатися в тому, що переліченим нижче характеристикам були надані оцінки n/a, 2, 3 або 4.**  9. Норми  10. Загальна надійність  11. Загальна валідність  12. Комп’ютерні звіти  Якщо будь-яка з даних характеристик має оцінку 0 чи 1, інструмент, як правило, буде відноситися до 1го, 2го, чи 3го пункту Рекомендацій, або ж до пункту “Інше” (з належним поясненням). | 1 Необхідна подальша розробка. Підходить лише для досліджень, але не для практичного застосування. | [ ] |
| 2 Підходить лише для використання експертом (з кваліфікацією користувача тестів ЄФПА вище 2го рівня) в ретельно контрольованих умовах або ж в дуже обмеженому колі областей застосування. | [ ] |
| 3 Підходить для використання під наглядом в сферах застосування визначених дистриб’ютором будь-яким користувачем тестів з загальною компетенцією в використанні та проведенні тесту (з кваліфікацією користувача тестів ЄФПА вище 2го рівня) | [ ] |
| 4 Підходить для використання в сферах застосування, визначених дистриб’ютором, користувачами тестів з компетенцією, яка вимагається дистриб’ютором (з якнайменш 2им рівнем кваліфікації користувача тестів ЄФПА) | [ ] |
| 5 Підходить для самооцінки без нагляду в сферах застосування, визначених дистриб’ютором | [ ] |
| 6 Інше …………………………………. | [ ] |

**ЧАСТИНА 3: БІБЛІОГРАФІЯ**

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment, 12*, 62–71.

Bartram, D. (2002a). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Evaluation Form*. Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D. (2002b). *EFPA Review Model for the description and evaluation of psychological instruments: Version 3.2. Notes for Reviewers.* Brussels: EFPA Standing Committee on Tests and Testing (September, 2002).

Bartram, D., & Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet*. Chichester, UK: Wiley and Sons.

Bartram, D., Lindley, P. A., & Foster, J. M. (1990). *A review of psychometric tests for assessment in vocational training.* Sheffield, UK: The Training Agency.

Bartram, D., Lindley, P. A., & Foster, J. M. (1992). *Review of psychometric tests for assessment in vocational training.* BPS Books: Leicester.

Bechger, T., Hemker, B., & Maris, G. (2009). *Over het gebruik van continue normering* [On the use of continuous norming]. Arnhem, The Netherlands: Cito.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. In D. Bartram & R. K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 201-217). Chichester, UK: Wiley and Sons.

Brennan, R. L. (Ed.) (2006). *Educational measurement*. Westport, CT: ACE/Praeger.

Cronbach, L. J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Hillsdale, NJ: Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (pp. 471-515).Westport, CT: ACE/Praeger.

Drenth, P. J. D., & Sijtsma, K. (2006). *Testtheorie. Inleiding in de theorie van de psychologische test en zijn toepassingen* (4e herziene druk) [Test theory. Introduction in the theory and application of psychological tests (4th revised ed.)]. Houten, The Netherlands: Bohn Stafleu van Loghum.

Embretson, S. E. (Ed.) (2010). Measuring psychological constructs. Advances in model-based approaches. Washington, D. C.: American Psychological Association.

Embretson, S. E., & Reise, S. P. (2000). Item response theory for psychologists.Mahwah, NJ: Erlbaum.

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of test ratings. International Journal of Testing, 1, 137–153.

Evers, A. (2001b). The revised Dutch rating system for test quality. International Journal of Testing, 1, 155–182.

Evers, A., Braak, M., Frima, R., & van Vliet-Mulder, J. C. (2009-2012). Documentatie van Tests en Testresearch in Nederland [Documentation of Tests and Testresearch in The Netherlands]. Amsterdam: Boom test uitgevers.

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2010). COTAN Beoordelingssysteem voor de Kwaliteit van Tests (geheel herziene versie; gewijzigde herdruk) [COTAN Rating system for test quality (completely revised edition; revised reprint)]. Amsterdam: NIP.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., et al. (2012). Testing practices in the 21st Century: Developments and European psychologists’ opinions. European Psychologist, in press.

Evers, A., Sijtsma, K., Lucassen, W., & Meijer, R. R. (2010). The Dutch review process for evaluating the quality of psychological tests: History, procedure and results. International Journal of Testing, 10, 295-317.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15, 309-334.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. Applied Psychological Measurement, 24, 355–366.

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

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. American Psychologist, 58, 78-80. International Test Commission. (2005).

International Guidelines on Computer-Based and Internet Delivered Testing. Bruxelles, Belgium: Author.

Kersting, M. (2008). DIN Screen, Version 2. Leitfaden zur Kontrolle und Optimierung der Qualität von Verfahren und deren Einsatz bei beruflichen Eignungsbeurteilungen [DIN Screen, Version 2. Guide line for monitoring and optimizing the quality of instruments and their application in proficiency assessment procedures.]. In M. Kersting. Qualitätssicherung in der Diagnostik und Personalauswahl - der DIN Ansatz (S. 141-210) [Guaranteeing quality in diagnostics and personnel selection (p. 141-210)]. Göttingen: Hogrefe.

Lindley, P. A. (2009). Reviewing translated and adapted tests: Notes and checklist for reviewers:5 May 2009. Leicester, UK: British Psychological Society. Retrieved from http://www.efpa. eu/professionaldevelopment/tests-and-testing.

Lindley, P.A. (2009, July). Using EFPA Criteria as a common standard to review tests and instruments in different countries. In D.Bartram (Chair), National approaches to test quality assurance. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Lindley, P., Bartram, D., & Kennedy, N. (2004). EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.3. Leicester, UK: British Psychological Society (November, 2004).

Lindley, P., Bartram, D., & Kennedy, N. (2005). EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.41. Brussels: EFPA Standing Committee on Tests and Testing (August, 2005).

Lindley, P., Bartram, D., & Kennedy, N. (2008). EFPA Review Model for the description and evaluation of psychological tests: test review form and notes for reviewers: Version 3.42. Brussels: EFPA Standing Committee on Tests and Testing (September, 2008).

Lindley, P. A. (Senior Editor), Cooper, J., Robertson, I., Smith, M., & Waters, S. (Consulting Editors). (2001). Review of personality assessment instruments (Level B) for use in occupational settings. 2nd Edition. Leicester, UK: BPS Books.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., et al. (2001). Psychological testing and psychological assessment: A review of evidence and issues. American Psychologist, 56, 128-165.

Mokken, R. J. (1971). A theory and procedure of scale analysis. The Hague: Mouton.

Moosbrugger, H., Kelava, A., Hagemeister, C., Kersting, M., Lang, F., Reimann, G., et al. (2009, July). The German Test Review System (TBS-TK) and first experiences. In D. Bartram (Chair), National approaches to test quality assurance. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice ítems. Methodology, 2, 65-72.

Muñiz, J., & Bartram, D. (2007). Improving international tests and testing. European Psychologist, 12, 206-219.

Nielsen, S. L. (2009, July). Test certification through DNV in Norway. In D. Bartram (Chair), National approaches to test quality assurance. Symposium conducted at The 11th European Congress of Psychology, Oslo, Norway.

Nunnally, J. C., & Bernstein, I. H. (1994). Psychometric theory (3rd ed.). New York: McGraw-Hill.

Parshall, C. G., Spray, J. A., Davey, T., & Kalohn, J. (2001). Practical Considerations in Computer-based Testing. New York: Springer Verlag.

Prieto, G., & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España [A model for the evaluation of test quality in Spain]. Papeles del Psicólogo, 77, 65–71.

Reise, S. P., & Havilund, M. G. (2005). Item response theory and the measurement of clinical change. Journal of Personality Measurement, 84, 228-238.

Tideman, E. (2007). Psychological tests and testing in Sweden. Testing International, 17(June), 5–7.

Schneider, R. J., & Hough, L. M. (1995). Personality and industrial/organizational psychology. In C. L. Cooper & I. T. Robertson (Eds.), International Review of Industrial and Organizational Psychology, 10, 75-129.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. Statistical Methods in Medical Research, 7, 301-317.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. Science, 240, 1285-1293.

Testkuratorium. (2009). TBS-TK. Testbeurteilungssystem des Testkuratoriums der Föderation Deutscher Psychologenvereinigungen. Revidierte Fassung vom 09. September 2009 [TBS-TK. Test review system of the board of testing of the Federation of German psychologists’ associations]. Report Psychologie, 34, 470-478.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), Adapting educational and psychological tests for cross-cultural assessment. Mahwah, NJ: Erlbaum.

Van der Linden, W. J., & Glas, C. A. W. (Eds.) (2010). Elements of adaptive testing. London: Springer.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Erlbaum.

Ziegler, M., MacCann, C., & Roberts, R. (Eds.) (2011). New perspectives on faking in personality assessment. Oxford, UK: Oxford University Press.

**ДОДАТОК**

Пам’ятка ключових пунктів для коментування у випадку, коли інструмент було перекладено та/або адаптовано з нелокального контексту.

|  |  |
| --- | --- |
| **Розробка** | |
| Докази | *Внеску з боку носіїв нової мови* |
| *Неодноразової експертизи (чи рецензії) лінгвістів та контент-експертів* |
| *Проведення зворотного перекладу з нової мови на оригінальну* |
| Основні психометричні якості | *Функціональність питань* |
| *Надійність* |
| **Норми** | |
|  | *Наявна локальна норма* |
| Нелокальна норма | *Вагомий доказ еквівалентності версій тесту та вибірок* |
| Міжнародні норми | *Більші ніж звичайно вимагається від локальних вибірок* |
|
| Характер вибірки | *Баланс джерел вибірки* |
| *Еквівалентність вихідних даних різних частин вибірки* |
| Еквівалентність тестових версій | *Усі іншомовні версії якісно перекладено/адаптовано* |
| *Деякі групи склали перекладену/адаптовану версію тесту* |
| Схожість показників в різних вибірках | *Якщо є значні розбіжності, про це має бути зазначено, а висновки щодо використання обговорені* |
| Вказівки щодо генералізації норм |  |
| **Еквівалентність/ Надійність/ Валідність** | |
| Інваріантність структури конструкту | *У вигляді структури фактору, еквівалентності кореляційних матриць, або схожості шаблонів кореляції зі стандартними показниками* |
| Схожість критеріальної валідності | *Найсильніші кореляції зі схожими компетенціями* |
| Схожі шаблони? Наповнення завданнями | *Завдання (чи питання) корелюють з іншими шкалами за одною моделлю* |
| *Найсильніше/найслабше наповнення завданнями схоже в оригінальному і в перекладеному варіанті* |
| Надійність альтернативних форм | *Учасники тесту, що володіють двома мовами, мають схожі профілі в обох мовах* |
| **Генералізація валідності** | |
| Генералізація валідності потребує вагомих доказів | *При здійсненні перекладу тестів між різними мовними сім’ями (наприклад, з індоєвропейських мов на мови семітської групи)* |
| Генералізація валідності може бути самостійно виявлена | *Генералізація валідності може бути отримана з доказів інваріантності валідності в попередніх перекладах, якщо тест було перекладено на декілька мов. Наприклад, якщо тест на шведській мові було перекладено на французьку, німецьку та італійську мови, і еквівалентність цих варіантів тесту було доказано.* |

1. Кожна країна може зробити вибір: зазначити імена рецензентів при публікації інтегрованої рецензії, чи зробити її анонімною. [↑](#footnote-ref-1)
2. Ця інформація повинна надаватися редактором або адміністрацією. [↑](#footnote-ref-2)