Тєрєхов А. М., *старший науковий співробітник наукового центру мовного тестування навчально-наукового центру іноземних мов НУОУ (м. Київ)*
Жуков В. Є., *начальник науково-дослідного відділу -заступник начальника наукового центру мовного тестування навчально-наукового центру іноземних мов НУОУ (м. Київ)*

ТЕОРЕТИЧНІ ОСНОВИ ОЦІНЮВАННЯ ЯКОСТІ ЛІНГВОДИДАКТИЧНИХ ТЕСТІВ

У статті розглядаються проблеми оцінювання якості лінгводидактичних тестів на основі класичної теорії тестів та Item Response Theory (ключові слова – педагогічні тести, класична теорія тестів, Item Response Theory, IRT).

В статье рассматриваются проблемы оценивания качества лингводидактических тестов на основе классической теории тестов и Item Response Theory (IRT) (ключевые слова – педагогические тесты, классическая теория тестов, Item Response Theory, IRT).

Problems of assessing of didactic-linguistic tests quality on the basis of the Classical Item Analysis and Item Response Theory (IRT) are considered in the article (keywords – pedagogical tests, Classical Item Analysis, Item Response Theory, IRT).

Одним з проблемних питань системи мовної підготовки особового складу Збройних Сил України, що значно знижує її ефективність, є недостатня стандартизація мовних (лінгводидактичних) тестів [1]. Необхідність забезпечення об'єктивності та надійності результатів мовного тестування особового складу Збройних Сил України зумовлює актуальність удосконалення методики оцінювання якості іншомовних тестів, які розробляються та використовуються для визначення рівня іншомовної компетенції особового складу Збройних Сил України.

Поряд з основним завданням процесу тестування (визначення рівня знань, умінь та навиків) великого значення набуває проблема якісного укладання самих тестів. Йдеться про визначення і подальшу розробку концепцій, призначення, змісту, параметрів і характеристик тестових завдань, а також вимог і рекомендацій до них, послідовне й чітке дотримання котрих дозволило б укладати високоякісні тести як гарантію об'єктивного оцінювання результатів навчання та прищеплення фахових навичок. Саме тому на чільне місце виступає проблема забезпечення валідності та надійності тестових завдань як основних параметрів якості тесту.

Відправні положення щодо розроблення валідних та надійних педагогічних тестів, критеріїв навчальних досягнень та рівнів володіння іноземною мовою надають Загальноєвропейські Рекомендації з мовної освіти: вивчення, викладання, оцінювання [2]. Наголошуючи на необхідності врахування трьох "фундаментальних чинників контролю": валідності, надійності і достовірності, Загальноєвропейські Рекомендації з мовної освіти ставлять за мету лише "розробити довідкові положення, а не практичний інструмент контролю". Тому рекомендації щодо визначення якості тестів і зокрема валідності у цьому документі носять загальний, концептуальний характер.

Але методи визначення основних якісних характеристик тесту суттєво відрізняються в роботах різних дослідників, і часто відбивають протилежні точки зору на проблему. Так, одні автори стверджують, що "визначення надійності й валідності тестів проводиться шляхом статистичної обробки результатів масового тестування в різних групах учасників тестування" [3], і навіть ще більш категорично: "Наукова тестологія і практика показують, що головним аргументом в оцінці якості тестів виступає статистика. Будь-які оцінки й судження, навіть фахівцями високого рівня, побудовані на особистих враженнях, неминуче будуть мати відбиток суб'єктивізму. Тестові завдання доводять свою придатність тільки статистично" [4]. Інші дослідники, навпаки, говорять про важливість, прийнятність, а іноді й переваги оцінки якості тестів фахівцями-експертами: "Придатність тесту визначається, в основному, шляхом якісного оцінювання із залученням експертів" [5]. Причинами цього вказуються складність

математичної формалізації процедур та розрахунків оцінки валідності та надійності і необхідність розробки спеціального програмного забезпечення для статистичного аналізу тестів. Зокрема, змістовну валідність найчастіше пропонується оцінювати за допомогою експертної комісії, а не математичних обчислень.

Таким чином, для створення та використання збалансованих методик, які б враховували переваги та недоліки як математичних, так й експертних методів визначення якісних характеристик тесту, розробникам педагогічних тестів, зокрема лінгводидактичних, необхідно у своїй діяльності дотримуватися теоретичних основ оцінювання якості педагогічних тестів, яки напрацьовані людством за майже 150-річну історію використання тестів для перевірки знань, вмінь, навиків, здібностей, розумового розвитку та інших характеристик особистості.

Метою цієї статті є розгляд проблем оцінювання якості лінгводидактичних тестів як експертними, так і математичними методами на основі класичної теорії тестів та більш сучасної теорії тестування – Item Response Theory (IRT), та надання рекомендацій щодо шляхів розв'язання визначених проблем для удосконалення методики мовного тестування особового складу Збройних Сил України.

Історично виділяють два основні підходи до оцінювання якості тестів математичними методами.

Перший з них набув широкого розвитку в рамках **класичної (емпірико-статистичної) теорії тестів,** яка ґрунтується на статистичних методах аналізу результатів тестування. Суттєвий внесок у розвиток класичної теорії тестів внесли такі вчені, як Чальз Спирмен (Charles Spearman), Льюїс Гуттман (Louis Guttman), Гарольд Гулліксен (Harold Gulliksen), Лінда Крокер та Джеймс Елджина (Crocker Linda, Algina James). З російських дослідників вперше опис цієї теорії дав В.С. Аванесов у 1989 р. У роботі Челишковой М.Б. [7] у 2002 р. наведено базову методику статистичного обґрунтування якості тесту.

Класична теорія тестів ґрунтується на таких основних положеннях [8].

1. Емпірично отриманий результат вимірювання X являє собою суму істинного результату вимірювання T і похибки виміру E:

$$X = T + E$$

Показники T і E зазвичай невідомі.

2. Істинний результат вимірювання можна виразити як математичне очікування $E(X)$:

$$T = E(X)$$

3. Кореляція істинних і помилкових компонентів дорівнює нулю.

4. Помилкові компоненти будь-яких двох тестів не корелюють між собою.

5. Помилкові компоненти одного тесту не корелюють з істинними компонентами будь-якого іншого тесту.

Рівень знань учасників тестування оцінюється за допомогою їх індивідуальних балів. Бал обчислюють як алгебраїчну суму оцінок виконання кожного завдання тесту. Результати тестування звичайно представляються у вигляді матриці з рядками та стовпцями, яка відображає результат виконання всіх завдань учасниками тестування. На практиці прийнято, як правило, використовувати дихотомічну шкалу оцінок результатів: у результаті правильного виконання завдання учасник тестування отримує один бал, в протилежному випадку – нуль балів. Первинна статистична обробка матриці результатів тестування має на меті визначити складність завдань на основі емпіричної перевірки завдань, як відношення кількості правильних відповідей до загальної кількості завдань у тесті. Визначення рівня складності тестового завдання і всього тесту є основоположним у визначенні результатів тестування, оскільки на цих показниках базуються подальші розрахунки валідності та надійності тестів.

У класичній теорії тестів багато років розглядалися тільки емпіричні показники складності. Проте емпіричний тестовий бал (X) залежить від багатьох умов – рівня складності завдань, рівня підготовленості випробуваних, кількості завдань, умов проведення тестування тощо. Таким чином, найважливішою проблемою класичної теорії тестів є визначення істинного тестового бала випробуваного (T). Для дослідження результатів виміру ця теорія використовує розвинутий класичний статистичний апарат. В його рамках можна проводити кореляційний аналіз результатів тестування, а також встановлювати кореляцію між якістю виконання конкретного завдання й тесту в цілому. За допомогою класичної теорії можна передбачити результати виконання тесту, виявити валідність окремих завдань тесту, обчислити надійність та валідність тесту в цілому як системи тестових завдань. Для цього розроблено та застосовуються багато методик, детально описаних у [7-11].

Разом з тим, штучність низки припущень класичної теорії тестів і деякі її практичні недоліки помітно вплинули на ріст критичних тенденцій. У першу чергу, цьому сприяли сумніви щодо правомірності традиційного оцінювання складності завдань за допомогою частки правильних чи неправильних відповідей. Адже при традиційному підході до зміни рівня складності завдань на різних за підготовкою вибірках учасників тесту залишається відкритим питання про об'єктивність значень параметра складності завдань тесту [9].

Спроба введення вагових коефіцієнтів, що відображають вклад окремого завдання в індивідуальний бал учасника тесту, суттєво не виправляє такі недоліки. Значення цих коефіцієнтів, у свою чергу, можна поставити під сумнів: деякі з них визначаються суб'єктивно, на основі думки педагога про складність завдання, решта базуються на емпіричних даних тестування і, відповідно, залежать від вибірки учасників тесту.

Таким чином, можна відзначити, що нестійкість статистик та їх взаємний вплив помітно знижують вірогідність тестових результатів. За допомогою цих статистик не можна об'єктивно оцінити значення параметрів, що характеризують складність завдання тесту, а також виразити значення цих параметрів на інтервальній шкалі [7].

Отже, класична теорія тестування, незважаючи на добре розроблений математичний апарат, піддається критиці за принципові недоліки. Зокрема, тестові бали учасників тестування залежать від складності завдань у тесті, **а складність завдання залежить від вибору учасників тестування**. Ще більшим недоліком класичної теорії вважається **нелінійність тестових балів** учасників тестування.

У нових варіантах педагогічних теорій тестів більше уваги стало приділятися характеру розумової діяльності учасників тестування у процесі виконання тестових завдань різних форм, тобто не емпіричним, а умоглядним способам оцінки підготовленості випробуваних. Так, в останні півсторіччя активно розвивається інший підхід до оцінювання якості тесту – більш сучасна теорія тестування Item Response Theory (IRT), що є частиною загальної теорії латентно-структурного аналізу. Її назва дослівно перекладається як "Теорія відповіді на завдання", але слід зауважити, що загальноприйнятої російсько-та україномовної назви для цієї теорії поки немає, тому у різних джерелах вона може називатися по-різному – "Теорія моделювання та параметризації педагогічних тестів", "Математико-статистична теорія оцінки латентних параметрів заданий тесту та рівня підготовленості тих, хто тестується", "Стохастична теорія тестів" тощо. Найчастіше використовують абревіатуру без перекладу – IRT.

Основні припущення IRT:

існують латентні (приховані) параметри особистості, недоступні для безпосереднього спостереження. У тестуванні це рівень підготовленості випробуваного й рівень складності завдання;

існують індикаторні змінні, пов'язані з латентними параметрами, доступні для безпосереднього спостереження. За значеннями індикаторних змінних можна судити про значення латентних параметрів;

оцінюваний латентний параметр повинен бути одномірним. Це означає, що тест, повинен вимірювати знання тільки в одній, чітко заданій, предметної області. Якщо умова одномірності не виконується, то необхідно переробити тест, видаливши завдання, що порушують його гомогенність.

Існують й інші припущення, що носять спеціальний характер і пов'язані з математико-статистичним апаратом IRT для обробки емпіричних даних.

Основним завданням IRT є перехід від індикаторних змінних до латентних параметрів, а основним змістом – вимір латентних параметрів, а саме рівня підготовленості тих, хто тестуються, і складності завдань тесту [8]. Головним в IRT є твердження про залежність ймовірності правильної відповіді того, хто тестується, від рівня його підготовленості та від параметрів завдань. Цю залежність зручно представляти у вигляді так званої логістичної функції. Число розглянутих параметрів ставиться у відповідність до моделі виміру. Модель виміру визначається як структурна побудова, що дозволяє з'єднати латентні змінні з одним або з більшим числом емпірично спостережуваних змінних.

Датський математик Г. Раш (G. Rasch) спробував формалізувати ідею залежності результатів тестування від співвідношення рівня підготовленості кожного випробуваного з мірою складності кожного завдання [12]. За його теорією Rasch measurement, основні латентні параметри, що впливають на результати тесту, – це рівень підготовленості випробуваних $\boldsymbol{\theta}$ і складність завдання тесту $\boldsymbol{\beta}$. Однак на практиці завжди ставиться зворотне завдання: за індикаторами

– відповідями на завдання тесту оцінити значення латентних параметрів θ і β. Для його рішення потрібно відповісти щонайменше на два питання. Перше – пов'язане з вибором виду співвідношень між латентними параметрами θ й β. Г. Раш запропонував увести його у вигляді єдиного параметру – різниці $(\theta - \beta)$, припускаючи, що параметри θ й β оцінюються за одною шкалою (тому теорією Rasch measurement ще називають однопараметричною теорією IRT). Якщо ця різниця позитивна і велика, то відповідно високою є ймовірність досягнення успіху *i*-го випробуваного в *j*-му завданні, і навпаки – якщо ця різниця негативна і велика за модулем, то ймовірність досягнення успіху *i*-го випробуваного в *j*-му завданні $P_j(\theta)$ буде низькою:

$$P_j(\theta) = \frac{e^{1,7(\theta-\beta_i)}}{1+e^{1,7(\theta-\beta_i)}} \qquad (1)$$

Відповідь на друге питання, що є центральним у теорії IRT, пов'язана з вибором математичної моделі для опису розглянутого зв'язку між латентними параметрами й спостережуваними результатами виконання тесту. Така модель визначається формулою, що задає умовну ймовірність правильної відповіді випробуваного на завдання тесту. Ця формула й визначає логістичну функцію. В однопараметричній моделі Г. Раша в якості логістичної функції вибирають модель, описану формулою 1, до очевидних переваг якої належить порівняльна стійкість значень рівня знань і складності завдання, що вираховуються. При цьому модель Раша вкрай незвична, вона суперечить стандартній парадигмі наукових досліджень. За класичним підходом – якщо теорія погано описує емпіричні дані, то її треба покращувати. За поглядами Г. Раша навпаки – якщо емпіричні дані суперечать його теорії, то ці дані слід відкинути, вони є недостовірними, внаслідок, наприклад, неточностей у формулюванні завдань, порушень в процедурі тестування тощо. Тому однопараметрична теорія IRT Г. Раша є прикладом іншої філософії виміру – model based measurement – вимірювань, що ґрунтуються на обраній моделі, де стверджується протилежне – не модель повинна відповідати емпіричним даним, а дані повинні відповідати моделі. Про це можна сперечатися, але згідно з цією філософією педагогічний тест утворюють тільки ті завдання, які відповідають даній моделі вимірювання. Всі інші завдання в тест не включаються.

При цьому однопараметрична модель Раша перетворює виміри, зроблені в дихотомічних і порядкових шкалах в лінійні виміри, в результаті якісні дані аналізуються за допомогою кількісних методів. Це дозволяє використовувати широкий спектр статистичних процедур. Крім того, модель Раша характеризується найменшим числом параметрів: один параметр рівня знань для кожного випробовуваного і тільки один параметр складності для кожного завдання. Аналіз результатів тестування на основі підходу Rasch measurement дозволяє оптимізувати зміст тесту і перетворювати його на інструмент для вимірювання рівня знань учасників тесту. Аналіз латентних параметрів та характеристичних кривих, побудованих за формулою 1 для завдань різних рівнів складності, дозволяє оцінити ймовірність правильного виконання завдань тесту учасниками тесту з будь-яким рівнем підготовленості у вибірці до проведення тесту, і, відповідно, формувати оптимальні для даного рівня підготовленості випробовуваних тести.

Але у літературі можна зустріти чимало критики з приводу незастосовності моделі Раша до безлічі "тестів", і тому ведеться пошук інших моделей, більш адекватних отриманим результатам. Наприклад, А. Бірнбаум (A.Birnbaum) з цією метою в своїй двохпараметрічній моделі [13] запропонував ввести до логістичної функції ще один параметр, що характеризує здатність завдання до диференціації учасників тесту, а саме α_j – параметр роздільної здатності *j*-го завдання:

$$P_j(\theta) = \frac{e^{1,7\alpha_j(\theta-\beta_i)}}{1+e^{1,7\alpha_j(\theta-\beta_i)}}$$

Для ще кращої відповідності емпіричним даним А.Бірнбаум створив трьохпараметрічну модель шляхом введення третього параметру c_j, який враховує ймовірність вгадування правильної відповіді на *j*-те завдання:

$$P_j(\theta) = c_j + (1-c_j)\frac{e^{1,7\alpha_j(\theta-\beta_i)}}{1+e^{1,7\alpha_j(\theta-\beta_i)}}$$

Але аналіз вірогідності отриманих результатів, тобто ступеню придатності моделей IRT для цілей вимірювання латентних параметрів, показує, що найкраще відповідає вимогам, що ставляться до якісного вимірювального інструментарію, саме однопараметричная модель Раша [8]. Крім того, якщо однопараметричная модель Раша для отримання достатньо вірогідних результатів статистичного аналізу потребує

вибірки обсягом біля 100 суб'єктів, то двохпараметрічна модель А. Бірнбаума – більше 200 суб'єктів, а його ж трьохпараметрічна модель – більше 1000 суб'єктів [10].

Результати порівняння класичної теорії тестування і IRT детально подані у [15]. Відповідно до них, IRT має певні переваги перед класичною теорією тестів:

IRT (особливо це відноситься до моделі Раша) перетворює вимірювання, виконані в дихотомічних і порядкових шкалах, у лінійні вимірювання, в результаті якісні дані аналізуються за допомогою кількісних методів;

міра вимірювання параметрів моделі Раша є лінійною, що дозволяє використовувати широкий спектр статистичних процедур для аналізу результатів вимірювань;

оцінка складності тестових завдань не залежить від вибірки досліджуваних, на яких вона була отримана;

оцінка рівня підготовленості тих, хто тестуються не залежить від набору тестових завдань, що використовується;

неповнота даних (пропуск деяких комбінацій "той, хто тестується, – тестове завдання") не є критичною.

Разом з тим IRT вимагає дуже великих обсягів статистичних розрахунків, причому в ітераційних циклах, та результатів тестування великої вибірки (від 100 до більше ніж 1000 суб'єктів). Без обчислювальної техніки та розробки спеціальних програмних продуктів її практичне використання неможливе. Крім того, недоліком IRT є ігнорування проблеми валідності, оскільки дуже рідко відповіді на завдання тестів залежать від одного фактора. Тому дану технологію доцільно застосовувати для простих гомогенних тестів.

У ході статистичної обробки результатів апробаційного тестування також обов'язковим є дистракторний аналіз. Статистична обробка даних може показати, що деякі дистрактори (невірні, "відволікаючі" варіанти відповіді) не працюють, тобто учасники тестування їх взагалі не обирають внаслідок очевидної помилковості, і тоді завдання, наприклад, із трьома відповідями перетворюється за суттю в завдання із двома відповідями. Внаслідок цього підвищується ймовірність вгадування правильної відповіді і знижується точність вимірювань. Для протидії цьому негативному явищу під час аналізу результатів пілотного тестування здійснюється підрахунок частоти вибору кожного дистрактору у всіх завданнях. Вважається, що кожен дистрактор повинен вибиратися не менш ніж 5 % учасників тестування. "Непрацюючі" дистрактори заміняють. Як відзначає В.С. Аванесов, "без дистракторного аналізу тестів не буває" [9].

Але попри всі позитивні сторони застосування математичних методів для оцінювання якості розробленого тесту деякі характеристики тесту доцільно досліджувати експертними методами.

На думку В.С. Аванесова існують два можливі способи визначення якості тестів: або умоглядно, теоретично (за допомогою експертів), або статистично – за допомогою математично-статистичного апарату [9]. Але практика показує, що для зменшення впливу на результат оцінки суб'єктивних факторів доцільно до експертних методів також залучати апарат математичної статистики.

Успіх створення тесту багато в чому залежить від якості тестового матеріалу, яке забезпечується правильним плануванням змісту в специфікації тесту і умінням розробника коректно реалізувати цей план при розробці завдань тесту. Розробка завдань супроводжується відображенням змісту навчальної дисципліни (наприклад, при розробленні тестів досягнень) або розгорнутого опису знань, вмінь та навиків та визначених комунікативних ситуацій (наприклад, при розробленні кваліфікаційних тестів) в змісті тесту. Розширення числа тем і розділів навчальної дисципліни (переліку знань, вмінь, навиків та комунікативних ситуацій) веде до збільшення довжини тесту, що раціонально тільки до певних розумних меж. Тому при створенні тесту ставиться завдання відобразити головне. Підвищенню повноти відображення, а також досягнення ряду інших необхідних характеристик сприяє експертиза якості змісту тесту [9].

За оцінку якості тесту як математичними, так й експертними методами виступає і П. Клайн. Так, на його думку, надійність тесту можна виразити через:

1) коефіцієнт міжрейтерської надійності (inter-rater reliability), який визначають як коефіцієнт кореляції між результатами оцінювання двома або кількома рейтерами (фахівцями-експертами з тестування) одного й того ж контингенту тих, хто бере участь у тестуванні, і який повинен в оптимальному варіанті бути в межах 0,80-0,90;

2) коефіцієнт внутрішньорейтерської надійності (intra-rater reliability), який визначається як коефіцієнт кореляції між результатами оцінювання одним і тим же рейтером двох послідовних тестувань

однакового контингенту учасників тестування паралельною формою того самого тесту;

3) коефіцієнт надійності тесту (test-retest reliability), який визначають як коефіцієнт кореляції між результатами двох послідовних тестувань одного і того ж контингенту учасників тестування шляхом використання того самого тесту [14].

Оцінка якості змісту тесту звичайно проводиться за визначеною методикою незалежними експертами, які не брали участь у розробці тесту. Як правило, число експертів становить не менше трьох осіб з кожного тесту. До експертизи залучаються найбільш досвідчені викладачі. Перед початком роботи кожен експерт повинен ознайомитися зі специфікацією тесту, що містить пояснення щодо його структури та змісту.

З огляду на складність структури тестових вимірників і сукупності критеріїв, що визначають їхню якість, експертизу тесту і його елементів необхідно проводити на основі системного підходу.

Систему комплексної експертизи якості тестових матеріалів можна представити як процес, що складається із декількох етапів [7]. **Попередня експертиза** якості тестових завдань і тестів починається з оцінювання якості специфікації й кодифікатора тесту. Це один з найважливіших етапів комплексної експертизи в силу того, що якості тестових матеріалів оцінюються експертами відповідно до тих характеристик, які заявлені в цих документах. На цьому етапі здійснюється також оцінювання тестових матеріалів на відповідність їхнім формальним вимогам. Експерти-фахівці з предмету проводять змістовний аналіз формулювань тестових завдань й оцінюють їхню коректність. Підсумками попередньої експертизи є перелік завдань у тестовій формі, що відповідають всім вимогам, і перелік завдань, не відповідних тим чи іншим вимогам. У першому випадку комплекти завдань відправляють на наступний етап комплексної експертизи – внутрішню експертизу, у другому – повертають авторам з коментарями щодо причин відбраковування.

Внутрішня експертиза якості тесту, у свою чергу, включає чотири етапи.

Перший етап являє собою багатокомпонентний аналіз змісту тестового завдання, а саме:

предметно-змістовний аналіз, тобто точність відображення фактичного матеріалу навчальної програми і коректність подання змісту у формулюванні тестового завдання;

композиційний аналіз: композиція тестового завдання повинна являти собою єдність форми, змісту, інструкції з виконання завдання та різноманітних допоміжних компонентів (такі як, таблиці, малюнки, графіки);

функціональний аналіз: від точності визначення функціональної значимості тестового завдання залежить ефективність його застосування в процедурі тестування;

вербальний аналіз – граматична побудова форми тестового судження. Від правильності граматичного оформлення всіх компонентів композиції залежить чіткість, логічність формулювання й однозначність сприйняття тестового завдання.

Слід мати на увазі, що зміст тесту визначається як оптимальне відображення вимог до рівня підготовки учасників тесту у системі завдань тесту [16]. Вимога оптимальності виділяє певні критерії якості відображення.

Перший критерій – повнота вимог до рівня підготовки учасників тесту кожним варіантом тесту. Чим повніше охоплення, тим вище змістовна валідність тесту, тим більше впевненість в обґрунтованості оцінок, отриманих учасниками тесту під час складання тесту. При оцінці за першим критерієм експерт повинен підтвердити або спростувати відсоток охоплення програми (вимог стандартів), заявлений автором у специфікації тесту. Отриманий експертом відсоток охоплення порівнюється з наведеним у специфікації тесту. Потім вираховується міра відхилення у вигляді різниці відсотків.

Другий критерій – правильність пропорцій змісту тесту. Необхідна також впевненість у тому, що завдання тесту охоплюють у правильній пропорції усі важливі аспекти дисципліни. Найчастіше при розробці тесту можливе зміщення пропорцій, так як тест легко перенаситити тими розділами змісту, за яким легше скласти завдання.

Третій критерій – перевірка відповідності змісту системи завдань до специфікації тесту. Ступінь невідповідності визначається підрахуванням відсотка завдань, не передбачених специфікацією за змістовним аспектом. Зіставлення запланованих у специфікації і реальних кількостей завдань у тесті проводиться шляхом обчислення різниць. Таким чином, підраховується загальна кількість завдань, які не відповідають специфікації тесту.

Другий етап – прогнозування успішності виконання даного завдання різними за рівнем

підготовки учасниками тестування. Це – одна з найбільш складних і трудомістких робіт експерта, що полягає у визначенні рівня складності завдання й орієнтовного часу, необхідного для його виконання.

Третій етап – оцінювання тесту в цілому за наступними критеріями:

відповідність тесту цілям навчання й тестування: наскільки розроблений тест здатний максимально точно діагностувати рівень підготовки тих, хто тестується, у відповідному виді контролю (вхідний, рубіжний або підсумковий);

практичність тесту, що полягає в доступності інструкцій і змісту завдань тесту для розуміння особами, які тестуються;

курикулярна, конструктна, функціональна, змістовна та критеріальна валідність тесту;

композиція тесту – оцінюється внутрішня узгодженість завдань у тесті залежно від його призначення, а також його гомогенності або гетерогенності. Основне завдання оцінювання внутрішньої узгодженості завдань у тесті – це перевірка сполучення підібраних завдань між собою, що повинно відображати структурну ієрархію моделі підготовки з навчальної дисципліни. При цьому експертові необхідно оцінити ефективність запропонованої розроблювачем схеми й способу розташування завдань у тесті. Поняття збалансованості містить у собі пропорційне наповнення тесту завданнями з різними рівнями складності. Таким чином, аналіз композиції тесту показує ступінь гармонійного подання ключових елементів змісту навчальної дисципліни й адекватність їхнього відображення в тесті.

Крім критеріїв, є загальні принципи, що сприяють в певній мірі правильному відбору змісту тестів. **Принцип репрезентативності** регламентує не тільки повноту відображення, але і значимість змістовних елементів тесту. Зміст завдань має бути таким, щоб з відповідей на них можна було зробити висновок про знання або незнання всієї програми розділу або курсу, що перевіряється. **Принцип системності** передбачає підбір змістовних елементів, що відповідають вимогам системності і пов'язані між собою загальною структурою знань. При дотриманні принципу системності тест можна використовувати для виявлення не тільки обсягу знань, а й для оцінки якості структури знань учасників тесту.

Четвертий етап – оформлення підсумків експертного оцінювання та рекомендацій щодо доопрацювання або корегування завдання. Експерт наводить своє загальне враження про зміст тесту.

Таким чином, на думку авторів для оцінювання якості лінгводидактичних тестів, що створюються для проведення мовного тестування особового складу Збройних Сил України, є доцільним поєднання експертних та математичних методів оцінки якості тесту під час різних етапів його розроблення. Під час фаз визначення концепції тесту, розроблення специфікації та плану тесту, складання переліку завдань та по звершенню розроблення тестових завдань необхідно проводити експертну оцінку за описаною системою комплексної експертизи. При цьому при наявності кількох експертів або експертних груп, які працюють паралельно незалежно один від одного доцільно також застосовувати апарат математичної статистики для усереднення отриманих від експертів результатів.

Після проведення апробації тестових завдань (пілотного тестування) обробку результатів пілотного тесту доцільно проводити математичними методами з використанням апарату класичної теорії тестування або IRT. Для розрахунку коефіцієнту критеріальної (емпіричної) валідності тесту як коефіцієнту кореляції Пірсона між отриманими емпіричними даними та зовнішнім критерієм у якості зовнішнього критерію доцільно використовувати незалежну експертну оцінку (прогноз) результатів. У якості експертів у таких випадках зазвичай залучаються викладачі, що здійснювали навчання у групах слухачів.

Такий підхід щодо постійного моніторингу якісних характеристик тесту у процесі його розроблення дозволить підвищити якість тестових матеріалів, що розробляються, та забезпечити об'єктивність та надійність результатів мовного тестування особового складу Збройних Сил України.

Література

1. Наказ Міністра оборони України від 01.06.2009 року № 267 "Про затвердження Концепції мовної підготовки особового складу Збройних Сил України та Плану реалізації концепції мовної підготовки особового складу Збройних Сил України". – К.: ГШ, 2009. – 42 с.

2. Загальноєвропейські Рекомендації з мовної освіти: вивчення, викладання, оцінювання. Переклад Шерстюк О., Київ, вид-во Ленвіт, 2003.

3. Вяткина Е. Современное состояние методов тестирования знаний за рубежом и в России // Инновации в образовании, 2004. Вып. 1.

4. Феськов Н. Еще раз о тестовых формах аттестации. Республиканский ин-т контроля знаний. Республика Беларусь, 2006, – 126 с.

5. Михайлычев Е.А. Дидактическая тестология // Народное образование, 2001.

6. Майоров А.Н. – Теория и практика создания тестов для системы образования. – М.: «Интеллект-центр», 2001. -296 с.

7. Челышкова М.Б. Теория и практика конструирования педагогических тестов: уч. пособие / М. Челышкова – М. : Логос, 2002. – 432 с.

8. Ким В.С. Тестирование учебных достижений. Монография. Изд-во Уссурийского гос. пед. ин-та, Уссурийск, Россия. 2007. – 208с.

9. Аванесов В.С. Композиция тестовых заданий. / В. С. Аванесов. – М : Центр тестирования Минобразования РФ, 2002. – 239 с.

10. Alderson J., Clapham C., Wall D. Language Test Construction and Evaluation. / J. Charles Alderson, Caroline Clapham and Dianne Wall – Cambridge, University Press, 1995. – 310 с.

11. Hughes A. Testing for Language Teachers / A. Hughes – Cambridge, University Press, 2005. – 251 с.

12. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen, Danish Institute of Educational Research, 1960. (Expanded edition, Chicago, The University of Chicago Press, 1980).

13. Birnbaum A. Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F.M. Lord and M.R.Novick. Statistical Theories of Mental Test Scores. Readinf Mass.: Addison-Wesly, 1968. – 479 с.

14. Клайн П. Справочное руководство по конструированию тестов. / Клайн П., пер. с англ. Е. Савченко. – М. : ПАН Лтд., 1994. – 283 с.

15. Федорук П.І. Використання адаптивних тестів в інтелектуальних системах контролю знань / П.І. Федорук // Прикарпатський національний університет ім. Василя Стефаника, м. Івано-Франківськ. Штучний інтелект. – 2008. – № 3. – С. 380-387.

16. Grotjahn, Rüdiger (2000): "Determinanten der Schwierigkeit von Leseverstehensaufgaben: Theoretische Grundlagen und Konsequenzen für die Entwicklung des TestDaF". In: Bolton, Sibylle (ed.): TestDaF: Grundlagen für die Entwicklung eines neuen Sprachtests. – C. 7-55.