результаты сравнительного анализа среднего балла ВНО абитуриентов и уровней интеллектуального развития и логичности умозаключений студентов-первокурсников. Автором сделан вывод, что ВНО не может быть инструментом качественного отбора контингента студентов, а может быть лишь частью механизма обеспечения качества подготовки абитуриентов.

Ключевые слова: *обеспечение качества, ВНО (внешнее независимое оценивание), оценивание, ВОУ (высшее образовательное учреждение).*

Lynova I. O. External testing as instrument of higher educational institutions qualitatively contingent of students.

The article is devoted to the external evaluation as an instrument of higher educational institutions qualitatively contingent of students. Also it presents the results of a comparative study of the External testing average score graduates and levels of intellectual development and logical reasoning first-year students. The author concluded that the External testing cannot be an instrument of qualitative selection of the contingent of students, and can only be the part of a mechanism ensuring quality of preparation applicants.

Keywords: *Quality Assurance, Independent External testing (ZNO), assessment, High School.*

Лісова Т. В., Миляник А. І.
Ніжинський державний університет імені Миколи Гоголя,
Львівський регіональний центр оцінювання якості освіти
(Ніжин, Львів, Україна)

ПЕРСПЕКТИВИ ВИКОРИСТАННЯ МАТЕМАТИЧНИХ МОДЕЛЕЙ ТЕСТІВ ДЛЯ ШКАЛЮВАННЯ РЕЗУЛЬТАТІВ ЗНО

У статті за даними пробного інтернет-тестування, за змістом та формою максимально наближеного до реального, розглянуто можливість представлення результатів ЗНО на єдиній інтервальній шкалі для об'єктивного вимірювання рівнів підготовленості учасників тестування та складностей завдань за допомогою математичної моделі Partial Credit.

Ключові слова: *модель Partial Credit, шкалювання, латентна характеристика, рівень підготовки учасника тестування, складність завдання.*

Поряд із звітом про проведення зовнішнього незалежного оцінювання у 2013 році на сайті Українського центру оцінювання якості освіти розміщено застереження щодо неможливості використання звітних матеріалів (регіональних даних) для складання рейтингів загальноосвітніх навчальних закладів, яке останнім часом набуло неабиякої популярності. Результати ЗНО на сьогодні не можуть використовуватись і як надійне джерело моніторингу якості освіти, оскільки внаслідок використання методу еквіпроцентильної нормалізації при підрахунку тестових балів, цей вид оцінювання є повністю нормо-орієнтованим, а отже, окремі тестові бали не несуть жодної інформації про знання та вміння осіб, які отримали ці бали. Все ж громадськість не полишає мрія довідатись об'єктивну інформацію про стан освіти, про те, що повинні знати та вміти наші випускники і що ж насправді вони вміють та знають, який навчальний заклад може забезпечити якіснішу підготовку.

Використання даних ЗНО в більш повному обсязі з метою інформування громадськості про стан освіти, визначення змісту та основних напрямів освітньої політики, прогнозування динаміки та основних тенденцій її розвитку, вироблення науково обґрунтованих рекомендацій щодо прийняття ефективних управлінських рішень в галузі освіти було б можливе завдяки застосуванню методів та моделей сучасної теорії тестів, яка дозволяє проводити об'єктивні вимірювання, інваріантні щодо контингенту учасників тестування та набору тестових завдань, на інтервальній шкалі, кожна точка якої може бути

поєднана із змістом тестів.

У роботах Ракова С. А., Соколова О. Ю., Гороха В. П. [1, 3, 4] достатньо повно обґрунтовано, що одним із засобів реалізації принципів об'єктивності оцінювання та порівнюваності результатів ЗНО різних сесій і різних років з кожного предмета ЗНО є застосування методу еквіпроцентильної нормалізації для шкалювання тестових балів учасників тестування з метою забезпечення їх вступу до ВНЗ. Зазначено, що використання даного методу є виправданим, оскільки дотримуються дві основні вимоги: 1) для тестування використовуються тести, що вимірюють один і той самий конструкт (у випадку тестів ЗНО – рівень навчальних досягнень з певного предмета); 2) вибірки учасників різних сесій тестувань з одного предмета статистично не розрізняються [4, с. 3]. При такому підході повністю реалізовується роль ЗНО, як механізму справедливого відбору абітурієнтів до ВНЗ, але залишається не використаною (або засекреченою) велика кількість інформації про реальний стан рівня освіти в Україні.

З моменту появи революційних робіт данського математика Г. Раша у кінці 50-х років минулого століття було проведено багато фундаментальних досліджень в галузі тестології, що привело до появи сучасної теорії тестування (IRT) та різних її узагальнень. До переліку робіт [6 – 8] уже класиків теорії тестування F. B. Baker, J. M. Linacre, G. N. Masters можна долучати величезну кількість праць, що з'являються понині у даній галузі. У Росії дослідження у галузі тестології та педагогічних вимірювань проводять Челишкова М. Б., Звонников В. І., Нейман Ю. М., Хлебніков В. А., Аванесов В. С., Маслак А. А., Ким В. С. та інші. У роботі [2] викладено основи сучасної теорії тестування, яка отримала у перекладі назву як теорія моделювання та параметризації педагогічних тестів, і яка з 2001 року використовувалась у Росії при шкалюванні результатів єдиного державного екзамену. Критиці такого підходу було присвячено багато робіт, [5] – одна з них. Дискусії щодо доцільності використання методів IRT для шкалювання результатів тестувань "високих ставок", якими є тести ЄДЕ чи ЗНО, як в Росії, так і в Україні не вщухають, але всі опоненти не виключають можливості паралельного використання методів IRT для поглибленого аналізу результатів з метою вдосконалення самого тесту та процедури проведення.

У статті розглядатиметься застосування математичних моделей сімейства Раша для обробки результатів ЗНО не стільки з метою пред'явлення результатів учасникам тестування, скільки з метою поглибленого аналізу системи тестових завдань, виявлення причин їх незадовільного функціонування, збору інформації про реальні досягнення учасників у термінах знань та умінь та їх відповідність державним вимогам і стандартам освіти. Вибір саме моделей сімейства Раша зумовлюється двома основними причинами: у рамках цих моделей оцінки латентних параметрів мають найменшу очікувану похибку та залежать лише від первинних балів, що не порушує принципу справедливості. Технічною підтримкою обрано програму Winsteps, у якій є можливість проводити всебічний аналіз результатів тестування у рамках усіх основних моделей сімейства Раша для тестів з дихотомічними та політомічними завданнями, які присутні у тестах ЗНО.

Для аналізу використовуються результати пробного інтернет-тестування з математики, що проводилось ЛРЦОЯО у 2012 році і за формою та змістом було максимально наближеним до реального. У тестуванні брало участь 157 учнів 10-х класів, з яких 96 – з великих міст та районних центрів, 52 – жителі сіл, 9 – не вказали місце проживання. Програма Winsteps дозволяє комбінувати різні моделі для дихотомічних та політомічних завдань у одному файлі. Для завдань з вибором однієї правильної відповіді з п'яти (1-20) та завдань відкритої форми (25-32) використовувалась модель Раша, а для завдань на встановлення відповідності (21-24) – модель Partial Credit з категоріями від 0 до 4, яка також належить до моделей сімейства Раша.

Необхідною умовою застосування більшості одномірних моделей сучасної теорії тестів є одномірність тесту та локальна незалежність завдань. Крім того, для моделей

Раша суттєвими є припущення про відсутність угадування учасниками тестування правильних відповідей на завдання та про однакову дискримінуючу здатність усіх завдань. Крім безпосередньої побудови оцінок параметрів моделей програма Winsteps дозволяє здійснювати перевірку виконання усіх припущень моделі та перевірку відповідності побудованої моделі емпіричним даним. У ході такої перевірки можна отримати багато корисної інформації для підтвердження надійності та валідності тесту. Розглянемо основні етапи аналізу результатів тестування у рамках вказаних моделей.

1. *Побудова оцінок параметрів моделі*. У Winsteps для побудови статистичних оцінок параметрів моделей використовується метод максимальної вірогідності з початковим наближенням за методом PROX. Отримано оцінки складностей завдань та рівнів підготовленості учасників тестування у логітах на інтервальній шкалі, що дозволяє їх порівнювати між собою у спосіб, неможливий для сирих балів чи процентильних оцінок. Результати учасників тестування у логітах (або у будь-якій іншій шкалі, отриманій далі лінійним перетворенням) є більш об'єктивними, при цьому зберігається ранжування учасників і не порушується основна роль ЗНО як засобу відбору до ВНЗ. Порівняння розподілів характеристик учасників та завдань на одній шкалі (рис. 1) дозволяє зробити висновок, що даний тест не забезпечує однакову точність вимірювання в усіх точках шкали, він легкий (діапазон варіювання складностей завдань від -2,09 до 1,64 логіти) і у ньому відсутні завдання для оцінювання сильних учасників (діапазон варіювання рівнів підготовки від -5,01 до 4,69 логіти). Тест вважається збалансованим і таким, що відповідає можливостям учасників тестування, якщо діапазон складностей завдань повністю перекриває діапазон рівнів підготовленості, розподіл учасників за рівнем підготовленості близький до нормального, а завдань за складністю – близький до рівномірного. Як бачимо, наш тест далекий від ідеального.

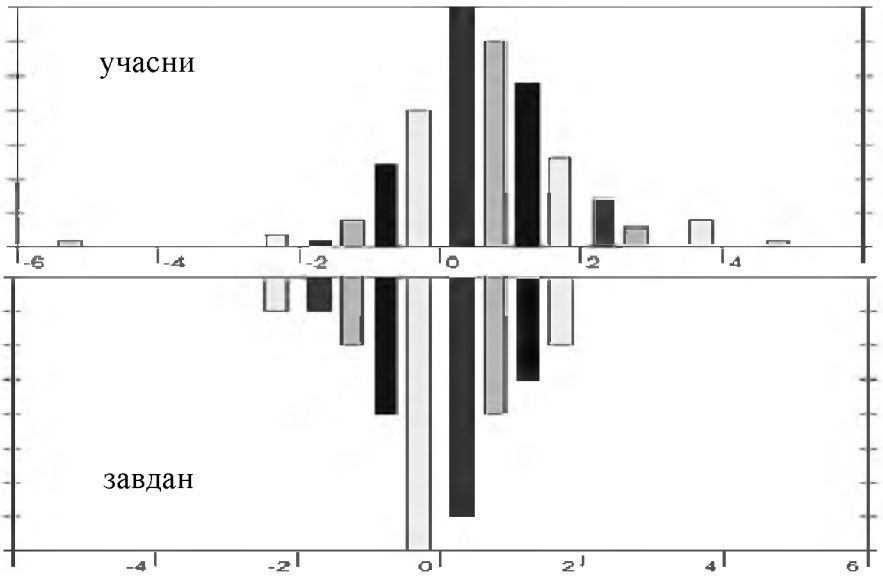

Рис. 1

2. *Розрахунок ймовірностей правильних відповідей на кожне завдання*. Відповідно до формул, які можна знайти в [2, 6, 8-9], розраховується ймовірність правильної відповіді (для політомічних завдань – очікуваний бал) для будь-якого учасника з континууму рівнів підготовки на кожне завдання тесту. Візуальна інформація подається у вигляді характеристичних кривих кожного завдання. Характеристичні криві якісного тесту повинні рівномірно (з кроком не менше 0,5 логіта) заповнювати проміжок від -5 до 5 логітів. На рис. 2 характеристичні криві перших 20 завдань з вибором однієї правильної

відповіді досить щільно покривають не широкий інтервал середніх та нижче середнього рівнів підготовки. Між двома найлегшими завданнями 6 та 1 є проміжок, на якому знаходяться ті рівні підготовки, для найкращого оцінювання яких у тесті відсутні завдання. Даний тест перевантажений завданнями однакової складності. Складнішим за 7 та 10 завдання є лише одне завдання 28 з відкритою відповіддю, складність якого, швидше за все, зумовлена неуважністю учасників, які переплутали поняття числа та цифри. Характеристичні функції завдань використовуються у процедурі Iitem Mapping, яка дозволяє кожну точку шкали прив'язати до змісту завдань тесту та знань і умінь, які цими завданнями перевіряються.

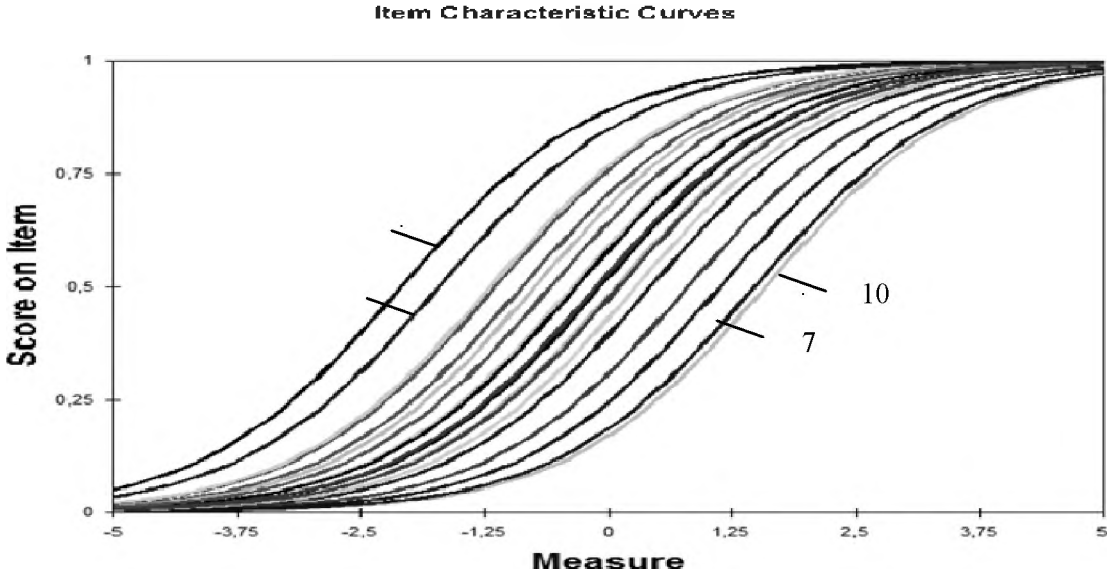

Рис. 2

Для політомічних завдань на встановлення відповідностей корисно аналізувати характеристичні криві категорій, за якими визначається ймовірність подолати певну категорію для кожного учасника. Так, для завдання 21 модель передбачає, що воно функціонує як дихотомічне з балами 0 або 4. Це означає, що визначення правильно однієї пари автоматично гарантує визначення всіх. Це недолік розробників даного завдання.

3. *Аналіз відповідності емпіричних даних побудованій моделі*. Статистичні процедури для перевірки відповідності емпіричних даних побудованій моделі у Winsteps реалізовані на основі стандартизованих залишків у вигляді незваженої (outfit), зваженої (infit) середньо-квадратичних статистик та їх стандартизованих значень. Високе значення outfit вказує на несподівані відповіді, що не узгоджуються з рівнем підготовки (вгадування, неуважність). Високе значення infit отримується тоді, коли несподівані відповіді на питання дають саме ті учасники, рівень підготовки яких відповідає складності питання. Це вказує на порушення валідності вимірювань і важче діагностується. Прийнятні значення середньо-квадратичних статистик знаходяться у межах від 0,5 до 1,5. Більші за 1,5 значення можуть вказувати на відхилення від одномірності в даних, менші за 0,5 – на явище overfit, коли дані настільки добре відповідають моделі, що у це важко повірити. Такі дані є менш продуктивними при вимірюванні, але не складають великої загрози для точності вимірювання. Розробники програми рекомендують проводити аналіз у такий спосіб, щоб спочатку виявити ті дані, що становлять більшу загрозу для валідності вимірювань: статистики outfit аналізуються перед infit, середньо-квадратичні статистики перед їх стандартизованими значеннями, великі значення перед малими.

Щоб мати впевненість у об'єктивності побудованої одномірної шкали для оцінювання конкретної змінної (а саме рівня підготовленості), дані, які не узгоджуються з

моделлю Раша, повинні бути вилучені з аналізу. І тут маємо парадокс – завдання з високою дискримінуючою здатністю підлягають вилученню. У [5] наводиться приклад, коли за перевіркою трьох випадково вибраних варіантів ЄДЕ забракувати довелось би від 50% до 70% завдань та не менше 10% учасників. Тому використання моделі Раша передбачає ретельну підготовку протягом кількох років з глибоким аналізом та апробацією завдань. У даному тесті маємо лише одне завдання 32 (outfit =1,69; infit =1,34), відповіді на яке швидше вгадувались і яке потребує незначного змістовного вдосконалення. Можна виділити не більше 10 учасників, які мали проблеми з розумінням завдань та отримали оцінку свого рівня підготовленості за рахунок вгадування.

4. *DTF та DIF аналіз*. Модель Раша дозволяє провести дослідження неупередженого функціонування тесту щодо різних підгруп опитаних та кількісно виявити різницю у вимірюваних характеристиках для двох або більше підгруп. При цьому використовуються два підходи: проводиться аналіз тесту як єдиного цілого, що складається з окремих питань, призначених для вимірювання однієї і тієї ж латентної характеристики (differential test functioning – DTF), або аналіз кожного завдання окремо з точки зору його функціонування в різних підгрупах (differential item functioning – DIF). Аналіз DTF можна проводити з метою дослідження одномірності тесту.

На рис. 3 наведено результат DTF аналізу для двох груп учасників: жителів міста та району. У цілому тест не виявляє упередженості щодо місця проживання, оскільки більшість завдань знаходяться у межах 95% довірчого інтервалу навколо лінії еквівалентності (пунктирна пряма) [7, с. 83], одне лише завдання 13 статистично значимо є складнішим для сільських учнів. Результат очікуваний, оскільки традиційно у непрофільних класах нерівності з модулями залишаються без належної уваги, що може бути предметом обговорення на семінарах вчителів району.

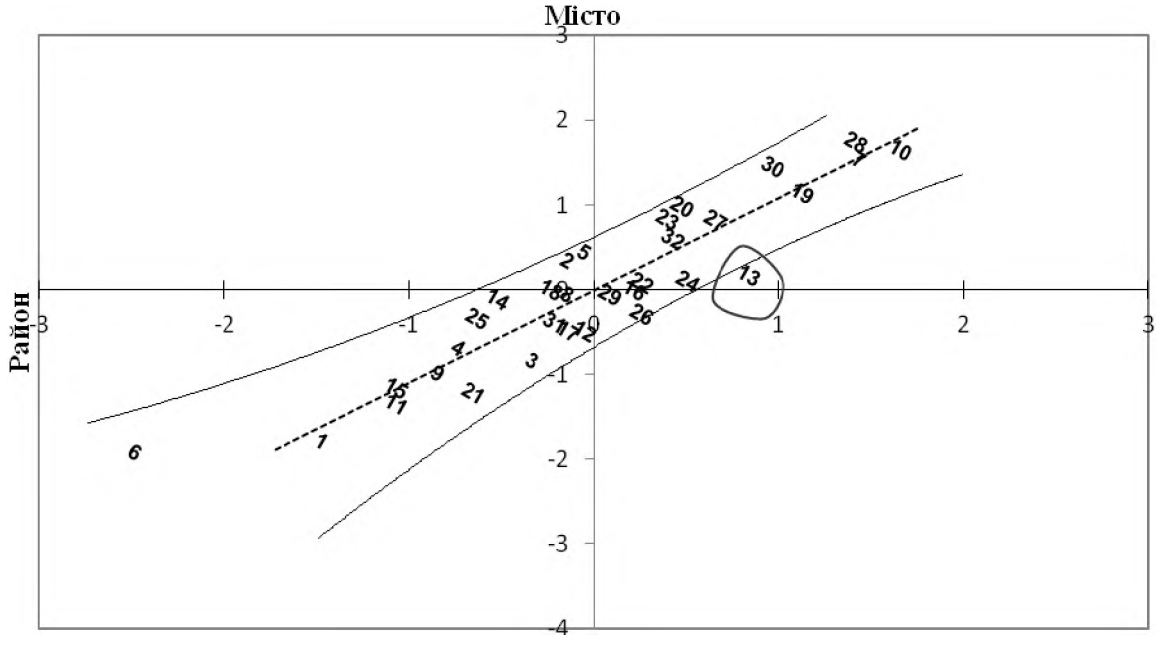

Рис. 3

Розрізняють різні прояви DIF: рівномірне DIF вказує на те, що міжгрупові відмінності зумовлені складністю питання, а нерівномірне DIF вказує на міжгрупові відмінності у дискримінуючій здатності питання. Наприклад, завдання 13 статистично значимо є складнішим для сільських учнів з середнім та нижчим за середній рівнем підготовки.

Програма Winsteps дозволяє проводити також DPF (differential person functioning)

аналіз для учасників тестування по відношенню до різних типів питань. Встановлено, що рівень підготовки вище середнього формується переважно за рахунок кращого виконання завдань з вибором однієї відповіді, учні міських шкіл краще (але статистично не значимо) впорались із завданнями на встановлення відповідності, для слабких учасників поганий результат переважно отримано внаслідок невдалого розв'язання завдань відкритого типу. При виявленні DIF чи DPF розробниками Winsteps рекомендуються різні варіанти дій залежно від їх величини та причини.

Висновки та перспективи подальших розробок дослідження проблеми. Тут розглянуто переважно ті аспекти аналізу результатів тестування ЗНО за допомогою математичних моделей, які можуть бути корисні при формуванні банку тестових завдань із заданими характеристиками. Щоб отримати інформацію про досягнення учасників тестування у термінах знань та умінь, необхідно далі провести процедуру Item Mapping для розміщення завдань на шкалі залежно від рівня RP (Response Probability) та дати змістову інтерпретацію кожної точки шкали (Scale Anchoring). Дана процедура була б корисною, поряд з іншими методами, для визначення порогових балів (Standard Setting) та для оцінки відповідності (Evaluate Alignment) між вимогами стандартів освіти та інструментами вимірювання.

Використана література:

1. *Горох В.* Порівняльний аналіз шкалювання результатів ЗНО різними методами / В. П. Горох, О. Ю. Соколов // Вісник ТІМО. – 2010. – № 12. – С. 22–29.
2. *Нейман Ю.* Введение в теорию моделирования и параметризации педагогических тестов / Ю. М. Нейман, В. А. Хлебников. – М. : Прометей, 2000. – 172 с.
3. *Раков С.* Теоретичні засади шкалювання результатів ЗНО методом еквіпроцентильної нормалізації / С. А. Раков // Вісник ТІМО. – 2010. – № 12. – С. 11–20.
4. *Раков С.* Оцінювання результатів ЗНО за шкалою 100-200 / С. А. Раков, О. Ю. Соколов // Історія в сучасній школі. – 2012. – № 5 (129). – С. 2–6.
5. *Челышкова М.* Шкалирование результатов единого госэкзамена: проблемы и перспективы / М. Б. Челышкова, А. Г. Шмелев // Вопросы образования. – 2004. – № 2. – С. 168–186.
6. *Baker F.* The Basics of Item Response Theory / F. B. Baker. – Portsmouth NH : Heinemann Educational Books, 1985. – 131 p.
7. *Linacre J.* A user's guide to Winsteps [Електронний ресурс] / John M. Linacre. – 2011. – Режим доступу : http: //www.winsteps.com. – Заголовок з екрану.
8. *Masters G.* A Rasch model for partial credit scoring / Geoff N. Masters // Psychometrika. – vol 47, № 2. June, 1982. – P. 150–174.
9. *Ostini R.* Polytomous item response theory models / R. Ostini, M. L. Nering. – Australia : Measured Progress, 2006. – 120 p.

Лисовая Т. В., Миляник А. И. Перспективы использования математических моделей тестов для шкалирования результатов ВНО.

В статье по данным пробного интернет-тестирования, по содержанию и форме максимально приближенного к реальному, рассмотрена возможность представления результатов ВНО на единой интервальной шкале для объективного измерения уровней подготовленности участников тестирования и сложности заданий с помощью математической модели Partial Credit.

Ключевые слова: *модель Partial Credit, шкалирование, латентная характеристика, уровень подготовленности участника тестирования, сложность задания.*

Lisova T., Mylyanyk A. Perspectives of using mathematical models for scaling of results of External Independent Testing.

The article deals with the opportunity of reporting the results of the EIT in a single interval scale for the objective measurement of person ability and item difficulty, according to Partial Credit model. For this purpose were used the data of online testing, which content and form is close to real conditions.

Keywords: *Partial Credit Model, scaling, latent characteristic, person ability, item difficulty.*