УДК 303.214.3

Лісова Тетяна Володимирівна
доцент, кандидат фізико-математичних наук, доцент кафедри прикладної математики, інформатики та освітніх вимірювань
Ніжинський державний університет імені Миколи Гоголя, м. Ніжин, Україна
tan-lisova@ndu.edu.ua

МЕТОДИ ТА ПРОГРАМНІ ЗАСОБИ ДОСЛІДЖЕННЯ ВАЛІДНОСТІ ТЕСТОВИХ РЕЗУЛЬТАТІВ ДЛЯ ГРУП ТЕСТОВАНИХ З ПЕВНИМИ ІНДИВІДУАЛЬНИМИ ОСОБЛИВОСТЯМИ

Анотація. Необхідною умовою наявності упередженого оцінювання за деяким тестом є диференційоване (відмінне) функціонування завдань цього тесту щодо різних груп учасників тестування. У даній статті викладено ідеї деяких статистичних методів виявлення диференційованого функціонування завдань, розроблених у рамках основних підходів до моделювання результатів тестування: за допомогою таблиць спряженості, регресійних моделей, багатовимірних моделей та моделей сучасної теорії тестів IRT (Item Response Theory). Тут розглядаються методи Мантель-Ханзеля, логістичної регресії, SIBTEST та метод відношення правдоподібності. Вказуються особливості й умови застосування кожного методу. Проведено огляд існуючих безкоштовних програмних засобів, у яких реалізовано дані методи. На прикладі реальних даних проведено порівняння цих методів. Вказано на доцільність використання кількох методів одночасно з метою зменшення ризику помилкових висновків.

Ключові слова: диференційоване функціонування завдання; рівномірне DIF; нерівномірне DIF; логістична регресія; метод Мантель-Ханзеля; тест відношення правдоподібності.

1. ВСТУП

Постановка проблеми. Одним із доказів валідності тесту, як інструменту вимірювання, є його неупереджене функціонування в різних групах, що гарантує справедливість оцінювання. Але інколи в тестових оцінках може існувати зміщення (англійською – *bias*), причини виникнення якого можуть бути пов’язані з тим, що результати тестування обумовлені не тільки вимірюваним конструктом, але й іншими факторами, які є сторонніми відносно цього конструкту (наприклад, належністю людини до певної групи – культурної, етнічної, соціальної, гендерної тощо). Термін *bias,* зазвичай, асоціюється з несправедливим, упередженим оцінюванням. Відомі випадки, коли результати тестування переглядались чи навіть скасовувались із значними матеріальними затратами через обґрунтовані претензії однієї з груп щодо упередженого їх оцінювання. Зважаючи на важливість даної проблеми, компанія ETS (Educational Testing Service), яка була і залишається лідером у справедливому оцінюванні, проводила у 1986 році міжнародну наукову конференцію, на якій було представлено різні підходи до виявлення *bias*. Лише співробітники ETS опублікували з тих пір більше 100 досліджень і звітів з даної проблематики [7].

У більшості випадків для виявлення *bias* використовують процедуру диференційованого функціонування завдання (Differential Item Functioning – DIF). У минулому терміни DIF і *bias* були взаємозамінними словами, але з 1988 року Holland P. і Thayer D. розрізнили ці два поняття. Уведення більш прийнятного терміну DIF дозволило також чітко розмежувати ситуації, коли завдання проявляє дійсно *bias* чи просто *impact*. Зараз вважаємо, що DIF «відбувається тоді, коли екзаменовані з різних груп (референтної і фокусної), які мають однакові рівні вимірюваного конструкту,

© Т.В. Лісова, 2015.

показують різні ймовірності успіху» [6]. Однак, DIF є необхідною, але не достатньою умовою *bias*. Завдання може демонструвати DIF через те, що різні групи дійсно мають різні рівні вимірюваної здатності і тому по-різному відповідають на завдання. У цьому випадку немає підстав вважати оцінювання несправедливим, тому говоримо про *impact*. Якщо ж екзаменовані однієї групи мають менше шансів відповісти правильно, ніж екзаменовані іншої групи, причому це зумовлено впливом факторів, які не стосуються мети тестування, то говоримо про *bias* завдання. Отже, якщо завдання не проявляє DIF, то воно функціонує неупереджено. Проте, якщо на початковому етапі статистичними методами було виявлено DIF, то для констатації *bias* завдання необхідно залучати експертів для проведення поглибленого аналізу і з'ясування можливих причин *bias*. DIF може бути рівномірним або нерівномірним залежно від того, чи є взаємодія між членством у групі й рівнем підготовки. Рівномірне DIF існує, коли такої взаємодії немає, тобто шанси на правильну відповідь для груп відрізняються на одну й ту ж величину для всіх рівнів підготовки. Натомість, нерівномірне DIF присутнє, коли така взаємодія проявляється.

Аналіз останніх досліджень і публікацій. З кінця 80-х років минулого століття було розроблено багато різних статистичних методів для виявлення DIF у рамках трьох основних підходів до моделювання результатів тестування: за допомогою таблиць спряженості або регресійних моделей, сучасної теорії тестів IRT (Item Response Theory) та багатовимірних моделей. Серед цих методів є параметричні і непараметричні, одні з них базуються на сирих балах, інші – на отриманих оцінках здатності, одні краще виявляють рівномірне DIF, інші – нерівномірне DIF. Серед цих методів немає універсального, кожен має певні переваги й недоліки. Але всі вони в основі мають ідею перевірки нульової статистичної гіпотези про відсутність DIF.

Простий у використанні й водночас потужний метод Мантель-Ханзеля вперше для дослідження DIF використали Holland P. і Thayer D. у 1988 році [13]. Метод базується на аналізі таблиць спряженості і, після відхилення нульової гіпотези про відсутність DIF, дозволяє класифікувати завдання за розміром DIF (велике, середнє та незначне DIF). Shealy R. і Stout W. (1993) запропонували метод Simultaneous Item Bias Test (SIBTEST) для визначення DIF у рамках багатовимірної моделі, коли весь простір латентної змінної вважається багатовимірним (θ, η), де θ – основна вимірювана здатність, а η – сторонні фактори [17]. Обидва ці методи добре працюють під час визначення рівномірного DIF, але мають низьку здатність виявляти нерівномірне DIF. Метод логістичної регресії, введений Swaminathan H. і Rogers H. у 1990 році як альтернатива процедурі Мантель-Ханзеля, дозволяє з однаковим успіхом виявляти рівномірне і нерівномірне DIF [20]. Таку ж властивість мають методи, засновані на використанні моделей IRT. Thissen D., Steinberg L. і Wainer H. (1993) використовували тест відношення правдоподібності для перевірки гіпотези про рівність параметрів розширеної і звуженої моделей [21]. Такий тест дозволяє глибше аналізувати причини DIF, що зумовлені не лише складністю завдання, а й роздільною здатністю і схильністю до угадування. Хоча методи IRT мають багато теоретичних переваг, вони, як і всі інші параметричні методи, засновані на модельних припущеннях, а тому висновки залежать від того, наскільки добре спостережувані дані відповідають моделі.

Загалом проблемі диференційованого функціонування завдань і тесту в цілому (Differential Test Functioning – DTF) присвячено багато робіт, одні з яких висвітлюють місце й роль DIF і DTF у дослідженні валідності інструменту вимірювання (Lord F. (1980), Angoff W. (1988), Hambleton R. (1991), Zumbo B. (1999)), інші пропонують і вдосконалюють різні статистичні процедури і засоби для дослідження DIF (Roussos L., Stout W. (1996), Bolt D. (2002), Bond T. (2003), Gierl M. та ін. (2004), Liu O. та ін. (2008), Terris R. (2008), Choi S. та ін. (2011); González A. та ін. (2011), Svetina D., Rutkowski L.

(2014), Wen Y. (2014) та багато інших). Багато робіт присвячено порівнянню різних методів на реальних чи змодельованих даних і проблемі узгодженості висновків, отриманих за різними процедурами. Деякі автори радять використовувати в одному дослідженні, особливо з реальними даними, кілька методів одночасно, оскільки відомі випадки, коли різні методи призводять до різних висновків щодо кількості завдань з проявом DIF і розміру DIF (Gierl M. та ін. (1999), Ercikan K. та ін. (2004), Yildirim H., Berberoglu G. (2009), Acar T., Kelecioglu H. (2010)).

За останні десятиліття з'явилось багато досліджень у психології, медицині та в освітній галузі, де досліджується вплив DIF на справедливість оцінювання і пропонуються шляхи зменшення такого впливу. Дані різних міжнародних обстежень залишаються постійно у центрі уваги дослідників різних країн. Так, Wilson M., Xie Y. (2008) на прикладі аналізу даних міжнародного дослідження PISA запропонували фасетне узагальнення процедури DIF. Виявленню упередженого оцінювання у міжнародному дослідженні PIRLS присвячена робота Sandilands D., Oliveri M. та ін. (2012). Різні прояви DIF у дослідженні предметних тестів знайшли відображення в роботах Kim M. (2001), Walker C., Beretvas N. (2006), Bao H. та ін. (2009), Salehi M., Tayebi A. (2012). Schmidt W. та ін. (1998), Klieme E., Baumert J. (2001) наголошували, що суттєвим джерелом DIF може бути багатовимірність тестів, що використовуються в широкомасштабних дослідженнях. У звіті TIMSS 2003 (Mullis I. та ін., 2004) наголошується, що взаємодія між завданнями чи когнітивними рівнями за гендерною ознакою може суттєво відрізнятися для різних країн через відмінності культур й освітніх систем. Hambleton R. (1993, 1994) звертав увагу на прояви DIF у групах, що є носіями різних мов, а причини цього вбачав не лише у проблемах перекладу, а й у відмінностях різних культурних традицій і програм підготовки. Додаткові джерела DIF у міжнародних дослідженнях розглядались у роботах Van der Vijver F., Tanzer N. (1998), Ercikan K. та ін. (1998, 2002), Sireci S., Berberoglu G. (2000), Yildirim H. (2006). У роботах [1; 2] проведено аналіз DIF у дослідженні деяких психологічних конструктів.

Мета статті. Використання статистичних методів для виявлення DIF у процесі валідизації інструменту вимірювання вимагає від дослідника розуміння технічної складності відповідних процедур. Це стоїть на заваді поширенню методів DIF аналізу серед дослідників, що не мають спеціальної математичної підготовки (психологів, освітян). Утім, такий аналіз потребує спеціалізованих програмних засобів для роботи з великими масивами даних, які часто є комерційними. З огляду на це й відповідно до концепції даного видання, **метою** статті є стислий виклад основних статистичних методів виявлення DIF, їх порівняння з використанням реальних даних і надання рекомендацій щодо вибору програмних засобів з вільним доступом для такого аналізу. У статті зацікавлений читач знайде відповіді на запитання, як виявити можливу упередженість в оцінюванні, але тут не йтиметься про причини, які до цього призводять.

2. МЕТОДИ ВИЗНАЧЕННЯ DIF

2.1. Метод Мантель-Ханзеля

Непараметричний метод Мантель-Ханзеля (далі MH) засновано на припущенні про рівність загальних шансів успіху в кожній групі (референтній і фокусній). Для виявлення DIF за статистикою MH у дихотомічних завданнях складається серія таблиць спряженості розміру 2×2, де кожна таблиця базується на даних про осіб, які отримали однакові бали за досліджуване завдання. Такі таблиці записують для кожного сирого бала *j*, але можна також поділити групи на страти за процентильними відношеннями, за

оцінками рівнів підготовленості у логітах, просто на рівні проміжки тощо. Якщо між групами немає відмінностей, то шанси на успіх вони мають однакові, а тому відношення шансів мало б бути близьке до 1. Це відношення шансів на всіх стратах обчислюється за формулою $\alpha_{MH} = \frac{\sum_j A_j D_j / T_j}{\sum_j B_j C_j / T_j}$. Якщо α дорівнює 1, то це означає, що шанси для фокусної групи отримати правильну відповідь такі ж, як і для референтної групи. Якщо це значення перевищує 1, то це означає, що члени референтної групи виконують дане завдання краще порівняно з членами фокусної групи. Навпаки, завдання з показником меншим за 1 означає, що референтна група виконує завдання гірше, ніж фокусна.

Таблиця 1.

Таблиця спряженості на одному рівні для дихотомічного завдання

Рівень (страта) j	Бал за завдання		Разом
	1	0	
Референтна група	A_j	B_j	N_{rj}
Фокусна група	C_j	D_j	N_{fj}
Разом	T_{1j}	T_{0j}	T_j

Отже, можна сформулювати нульову гіпотезу про відсутність відхилення у шансах на успіх в обох групах (відсутність DIF). Для перевірки даної гіпотези використовують статистику χ^2_{MH}, яка має розподіл χ^2 з одним ступенем волі:

$$\chi^2_{MH} = \frac{\left\{\left|\sum_j (A_j - E(A_j)\right| - 0.5\right\}^2}{\sum_j \mathrm{var}(A_j)}, \text{ де } E(A_j) = \frac{N_{rj} T_{1j}}{T_j}, \ \mathrm{var}(A_j) = \frac{N_{rj} N_{fj} T_{1j} T_{0j}}{T_j^2 (T_j - 1)}.$$

За величиною α_{MH} можна судити про силу (розмір) DIF, але вона має несиметричний розподіл, тому для зручності обчислюють $\ln\alpha$, а щоб краще узгодити з параметром, який використовує компанія ETS для класифікації завдань, обчислюють показник $\Delta_{MH} = -2.35\ln(\alpha_{MH})$. За значеннями цього показника завдання поділяються на класи відповідно до рівня прояву DIF: рівень А (незначне DIF), коли $|\Delta_{MH}| < 1$; рівень В (помірне DIF), коли $1 \le |\Delta_{MH}| < 1.5$; рівень С (велике DIF), коли $|\Delta_{MH}| \ge 1.5$. Як правило, стурбованість викликають завдання рівня С зі значним проявом DIF.

Метод МН має низку недоліків і переваг. Серед переваг варто вказати те, що метод МН не вимагає великих вибірок і зводиться до простих обчислень. Він легко поширюється на випадок політомічних завдань. Головним недоліком є те, що він не пристосований для виявлення нерівномірного DIF. Для різних страт навіть значні відхилення протилежних знаків можуть компенсуватися, завдяки чому значення критерію не потрапить у критичну область, де відхиляється гіпотеза про відсутність DIF. Тоді як тут може бути присутнє значне нерівномірне DIF.

2.2. Метод SIBTEST

Відповідно до іншого непараметричного методу SIBTEST (Simultaneous Item Bias Test) DIF розглядається як різниця між ймовірностями отримати правильну відповідь, яка виникає внаслідок того, що учасники з різних груп з однаковим рівнем основної

вимірюваної здатності θ мають різні оцінки сторонньої (вторинної) здатності η, яка впливає на їх відповіді. SIBTEST перевіряє нульову гіпотезу про відсутність DIF, коли різниця між ймовірностями правильної відповіді на досліджуване завдання (або набір завдань) для учасників фокусної і референтної груп з істинним балом T дорівнює нулеві: $B(T) = P_R(T) - P_F(T) = 0$. Метод дозволяє досліджувати кожне завдання окремо, але основна його перевага в тому, що можна виділяти набір завдань (suspect subtest), у яких підозрюється вплив стороннього фактора, і порівнювати їх з рештою завдань (matching subtest), у яких вплив відсутній.

Для кожного бала k цих решти завдань відповідний істинний бал для референтної і фокусної груп оцінюється за допомогою лінійної регресії. Shealy R. і Stout W. (1993) ввели корекцію регресії, щоб врахувати відмінності між Отже, незміщеною оцінкою $B(T)$ є зважена сума $\hat{B}_U = \sum_k p_k (\overline{Y}^*_{R_k} - \overline{Y}^*_{F_k})$, де p_k – частка екзаменованих фокальної групи, що мають k балів, $\overline{Y}^*_{R_k}$ та $\overline{Y}^*_{F_k}$ — скориговані середні бали для референтної і фокусної груп за досліджуване завдання на всіх рівнях. Для перевірки нульової гіпотези використовується статистика $SIB = \hat{B}_U / \sigma(\hat{B}_U)$, яка має стандартний нормальний розподіл [17]. Roussos L. і Stout W. (1996) запропонували класифікацію завдань залежно від розміру виявленого DIF: рівень A (незначне DIF), коли нульова гіпотеза про відсутність DIF відхиляється, і $\left|\hat{B}_U\right| < 0.059$; рівень B (помірне DIF), коли $0{,}059 \le \left|\hat{B}_U\right| < 0.088$; рівень C (велике DIF), коли $\left|\hat{B}_U\right| \ge 0.088$.

SIBTEST корисний для виявлення прихованих вторинних розмірностей, які часто є причиною DIF. Його можна використовувати як доповнення в проведенні розвідувального факторного аналізу, досліджуючи різні набори завдань. Він узагальнений для політомічних завдань, модифікований для виявлення нерівномірного DIF. Хоча даний метод розроблявся для моделювання багатовимірних даних, його так само успішно можна використовувати, коли дані є одновимірними.

2.3. Метод логістичної регресії

Логістичну регресію (LR) використовують для передбачення дихотомічної (бінарної) змінної за допомогою одного чи кількох кількісних чи якісних предикторів. У нашому випадку залежною змінною є відповіді на завдання, а предикторами можуть бути сирі бали (або значення рівнів підготовки), належність до певної групи, взаємодія цих факторів тощо. Для виявлення DIF будують кілька регресійних моделей, поступово їх ускладнюючи додаванням нових членів. Потім за допомогою критерію χ^2 ці моделі порівнюють, щоб виявити, наскільки суттєвим є вплив нового доданка. Це дозволяє, на відміну від інших методів, виявити як рівномірне, так і нерівномірне DIF.

Відповідно до загальної моделі логістичної регресії ймовірність того, що бінарна змінна u (бал за досліджуване завдання) набуде значення 1, має вигляд $P(u=1) = \dfrac{e^z}{1+e^z}$, де $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$, а $X_1, X_2, \ldots, X_k$ – незалежні змінні, які можуть визначати дану ймовірність. Модель логістичної регресії ще записують у вигляді $\ln \dfrac{P}{1-P} = \operatorname{logit} P = z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$.

У першу чергу, відповідь на завдання обумовлена рівнем підготовленості θ, тому найпростішою є модель $z = \beta_0 + \beta_1 \theta$. Це базова модель або ж Модель 1. Далі розглядається розширена, порівняно з попередньою, модель, яка враховує належність

до певної групи: $z = \beta_0 + \beta_1\theta + \beta_2 g$ (Модель 2). Тут категорійна змінна g вказує на належність до групи, часто приймають $g = 1$ для членів фокальної групи і $g = 0$ – для референтної. Для порівняння цих моделей можна використати, наприклад, тест G^2 (див. пункт 2.4). Якщо на заданому рівні значущості різниця між цими моделями буде статистично значущою, тобто β_2 суттєво відрізняється від нуля, то це свідчить про суттєвий вплив нового доданка (фактора належності до групи), а отже, про рівномірне DIF. Відповідь про наявність нерівномірного DIF можна отримати, розглядаючи наступну модель $z = \beta_0 + \beta_1\theta + \beta_2 g + \beta_3\theta g$ (Модель 3), у якій останній доданок характеризує взаємодію рівня підготовки з належністю до певної групи. Розширену Модель 3 можна аналогічно порівняти з Моделлю 2 або 1. Наразі, під час порівняння моделей 1 і 2, 2 і 3 статистика G^2 має розподіл χ^2 з одним ступенем волі, а під час порівняння моделей 1 і 3 – з двома ступенями волі. Порівнюючи Модель 2 і 3, ми приймаємо або відхиляємо нульову гіпотезу, що $\beta_3 = 0$. Якщо вона відхиляється, то є суттєвий вплив належності до групи у взаємодії з рівнем підготовки, що свідчить про нерівномірне DIF. За знаками коефіцієнтів регресії можна також визначити, яка з двох груп має перевагу, відповідаючи на дане завдання.

У проведенні регресійного аналізу важливим показником якості моделі є коефіцієнт детермінації R^2, який вказує на частку дисперсії залежної змінної, що зумовлена регресійною моделлю, у загальній дисперсії. Високе R^2 (близьке до 1) означає, що регресійна модель добре передбачає поведінку залежної змінної, а вплив сторонніх факторів незначний. На жаль, додавання нових незалежних змінних завжди веде до збільшення R^2, що не завжди означає покращення моделі. Однак R^2 використовується в інших статистиках для визначення доцільності введення до моделі нових змінних. Для логістичної регресії з цією метою використовують псевдо коефіцієнти детермінації Nagelkerke R^2 і Weighted-Least-Squares R^2. Перший частіше використовується, бо автоматично виводиться у звітних документах більшості статистичних пакетів (SPSS, SAS). Кожен з цих коефіцієнтів може бути використаний для обчислення величини $\Delta R^2 = R_A^2 - R_C^2$, яка використовується для порівняння розширеної і спрощеної моделей з метою визначення розміру DIF.

Gierl M., Rogers W. та Klinger D. (1999) запропонували класифікацію завдань за величиною DIF, яка допомагає зменшити помилку першого роду в прийнятті рішення: рівень A (незначне DIF), якщо $\Delta R^2 < 0.035$ незалежно від того, приймається чи відхиляється гіпотеза; рівень B (помірне DIF), якщо $0.035 \le \Delta R^2 < 0.070$ при відхиленні нульової гіпотези; рівень C (велике DIF), якщо $\Delta R^2 \ge 0.070$.

Геометрична ілюстрація методу можлива, якщо будувати окремо базові логістичні моделі у референтній $z_0 = \beta_{00} + \beta_{10}\theta$ і фокальній $z_1 = \beta_{01} + \beta_{11}\theta$ групах. Якщо $\beta_{00} = \beta_{01}$ та $\beta_{10} = \beta_{11}$, то регресійні прямі зливаються і DIF відсутнє, якщо $\beta_{00} \ne \beta_{01}$, але $\beta_{10} = \beta_{11}$, то прямі паралельні і маємо рівномірне DIF, якщо ж $\beta_{10} \ne \beta_{11}$, то прямі перетинаються, що свідчить про нерівномірне DIF. Причому, коефіцієнти цих моделей пов'язані з коефіцієнтами Моделі 3 співвідношеннями: $\beta_2 = \beta_{01} - \beta_{00}$ та $\beta_3 = \beta_{11} - \beta_{10}$.

Метод логістичної регресії не потребує великих обсягів вибірок і виконання модельних припущень щодо розподілу характеристик, як у параметричних методах. Хоча інколи, особливо у роботах прикладного змісту, його відносять до параметричних методів, аргументуючи тим, що тут є формули, що визначають модель. Метод LR може

використовувати як дискретні значення сирих балів, так і об'єктивні оцінки рівнів підготовленості на неперервній осі. Він дає можливість якісно дослідити рівномірний чи нерівномірний прояви DIF, легко застосовується до політомічних завдань (за наявності відповідного програмного забезпечення). Але варто мати на увазі, що помилка прийнятого рішення залежить не лише від обсягів вибірок, а й від співвідношення між розмірами груп.

2.4. Метод відношення правдоподібності

Моделі IRT дозволяють особливо зручно і наочно демонструвати, як працює завдання у різних групах. Якщо завдання не виявляє DIF, то його характеристичні криві, побудовані в одній системі відліку за даними від різних груп, повинні були б збігатися або ж відрізнятися статистично не значимо. Різницю між кривими можна оцінити різними способами. Raju N. (1988) пропонував оцінювати площу між ними. Інший підхід полягає у порівнянні параметрів цих кривих. Оскільки кожна крива визначається параметрами, кількість яких залежить від моделі, що використовується, то суттєве відхилення одного з параметрів для різних груп призведе до того, що криві не збігатимуться. Отже, щоб виявити DIF, необхідно перевірити нульову гіпотезу про відсутність різниці між параметрами завдання, оціненими у різних групах.

Параметричний метод відношення правдоподібності IRT-LR (повна назва Item Response Theory Likelihood Ratio Test) був запропонований Thissen D., Steinberg L. та Wainer H. (1988) для порівняння якості наближення двох різних моделей залежно від кількості параметрів, але його успішно використовують і в проведенні DIF аналізу. Для цього порівнюють дві моделі: розширену (augmented) модель A, у якій параметри досліджуваного завдання вважаються різними для двох груп, і спрощену (compact) модель C, у якій вважається, що параметри завдання однакові для обох груп.

Різницю між цими моделями оцінюють за допомогою статистики $G^2 = G_2^2 - G_1^2 = -2\ln L_C - (-2\ln L_A)$, де $\ln L$ – логарифм функції максимальної вірогідності відповідної моделі. Функція максимальної вірогідності дорівнює добутку ймовірностей отримання l балів ($l = \overline{0, m_j}$, але для дихотомічних завдань це 0 або 1) за кожне j-те ($j = \overline{1,k}$) завдання для кожного i-го ($i = \overline{1,n}$) учасника: $L = \prod_{i=1}^{n}\prod_{j=1}^{k}\prod_{l=0}^{m_j} p_{lj}(\theta_i)^{u_{lij}}$, де $p_{lj}(\theta_i)$ – ймовірність для учасника i отримати l балів за завдання j (визначається відповідно до моделі, що використовується), $u_{lij} = 1$, якщо учасник i отримав l балів за завдання j, та $u_{lij} = 0$ – якщо ні. Для великих вибірок, коли розподіл учасників відомий, використовують функцію маргінальної максимальної вірогідності $L = \int_{-\infty}^{+\infty}\prod_{j=1}^{k}\prod_{l=0}^{m_j} p_{lj}(\theta)^{u_{lj}}\, g(\theta)\, d\theta$, де $g(\theta)$ – відома щільність розподілу учасників (часто це щільність нормального розподілу). Відомо, що величина G^2 має розподіл χ^2 з числом ступенів волі, що дорівнює різниці між кількістю параметрів у моделях A і C [21].

Процедура IRT-LR аналізу DIF деякого завдання j має три кроки: 1) оцінюються параметри обраної моделі спільно в обох групах учасників так, що параметри всіх завдань в обох групах вважаються однаковими (модель C), і обчислюється величина G_2^2; 2) параметри завдання j (один або кілька) оцінюються окремо в кожній групі, а параметри всіх інших завдань вважаються однаковими для обох груп (модель A), та обчислюється величина G_1^2; 3) обчислюється спостережене значення G^2. Якщо воно

перевищує критичне за заданого рівня значущості α, то нульову гіпотезу про відсутність різниці між параметрами завдань, оцінених у різних групах, відхиляють. Це свідчить про наявність DIF у даному завданні. Для однопараметричної моделі для виявлення DIF, зумовленого різною складністю завдання у групах, таку трикрокову процедуру необхідно провести стільки разів, скільки є завдань, підозрілих щодо DIF. Під час використання дво- і трипараметричної моделей кількість необхідних порівнянь збільшується, але з'являється можливість виявити, який саме з параметрів є причиною DIF. Отже, є можливість виявити як рівномірне, так і нерівномірне DIF. Після відхилення нульової гіпотези можна класифікувати завдання за розміром DIF [3]: рівень А (незначне DIF), коли $3.84 < G^2 < 9.4$; рівень В (помірне DIF), коли $9.4 \le G^2 \le 41.9$; рівень С (велике DIF), коли $G^2 > 41.9$.

Зважаючи на те, що процедура IRT-LR для виявлення DIF є досить громіздкою порівняно з іншими методами і не завжди затрачені зусилля є виправданими, вона має низку переваг. По-перше, жоден із відомих методів не дає такого глибокого аналізу причин DIF. По-друге, через те, що методи IRT дозволяють будувати оцінки латентних параметрів незалежно від груп опитаних і навпаки, ця процедура може знайти застосування у комп'ютерному адаптивному тестуванні, де інші методи не застосовні. Основними недоліками даної процедури є вимога великих вибірок у обох групах і вимога відповідності реальних даних модельним припущенням.

3. РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

Після проведеного дослідження з педагогіки чи психології неминуче настає етап обробки його результатів. Якщо на цьому етапі виникне потреба дослідження DIF, то досліднику потрібно визначитись, чи буде він самостійно реалізовувати вище згадані методи, витрачаючи час і зусилля на виявлення, можливо, неістотних ефектів, чи скористається вже розробленими засобами, розуміючи суть методу й інтерпретацію отриманих результатів. На сьогодні розроблена достатня кількість програмних засобів, які допомагають проводити дослідження DIF. Відповідні модулі включено до всіх комерційних пакетів для аналізу результатів оцінювання, але ми зосередили увагу на пошуку програм з вільним доступом, яких також виявилося немало. Далі коротко розглянемо можливості деяких з них.

Так, програма EZDIF (Waller N., 1998) дозволяє аналізувати дихотомічні завдання за допомогою методів MH та LR. Вона створювалась для роботи в DOS, але без проблем працює в режимі командного рядка Windows. Потрібно лише вказати імена вихідного файлу і файлів з даними, описати структуру цих даних, вручну вказати поділ на страти за завданнями, регулюючи самостійно товщину страт. Аналіз відбувається у два етапи. Завдання, які на першому етапі демонструють значне DIF, на другому етапі вилучаються з порівняння. Вихідний текстовий документ містить значення всіх статистик за методом MH, класифікацію завдань, стандартну похибку розміру DIF і рівень значущості *p* (ймовірність справедливості нульової гіпотези). Результати аналізу за методом LR містять усі коефіцієнти регресії, стандартні похибки, відповідні статистики і рівні значущості *p*. Програма не містить ніяких обмежень на кількість завдань чи учасників. Її можна безкоштовно завантажити за посиланням http://www.psych.umn.edu/faculty/waller/downloads.htm.

Програма Easy-DIF (González A., Padilla J., Hidalgo M. та ін., 2011) має зручний інтерфейс і можливості графічної візуалізації отриманих результатів. Вона дозволяє аналізувати тести з дихотомічними і політомічними завданнями за методами MH і стандартизації, має можливості для виявлення нерівномірного DIF шляхом поділу

екзаменованих на сильну і слабку групи. Після завантаження файлу з даними, оформленого за вимогами програми, і вибору опцій для аналізу в одному з двох вікон з'являються всі статистики для кожного завдання (рис. 1). Тут же можна змінювати налаштування і бачити результат цих змін. Саме тому Easy-DIF корисна для вивчення даних методів. Вихідний текстовий документ містить також усю інформацію по кожному завданню, але він не дуже зручний для подальшої обробки. Програма не може аналізувати тести, які містять одночасно і дихотомічні, і політомічні завдання. Програму Easy-DIF можна отримати безкоштовно, звернувшись до розробників за електронною адресою, вказаною у їх роботі [11].

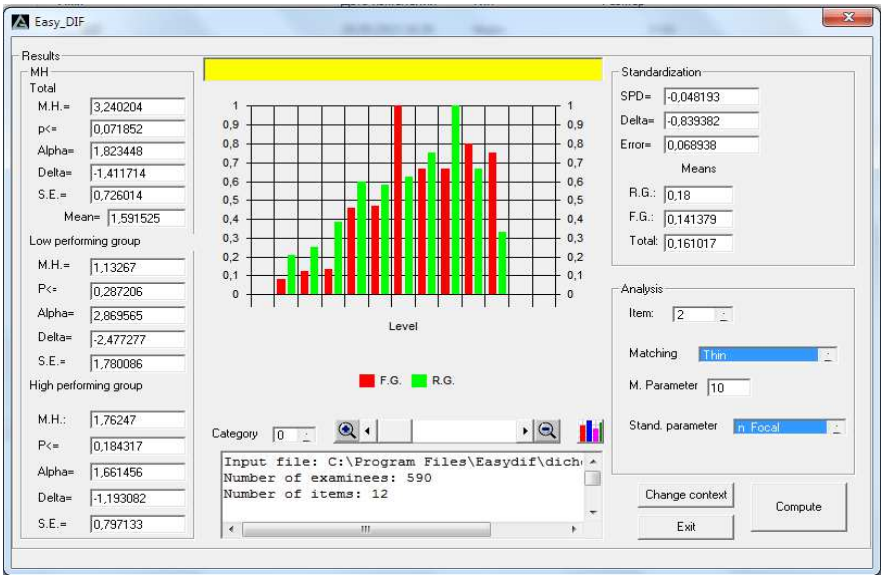
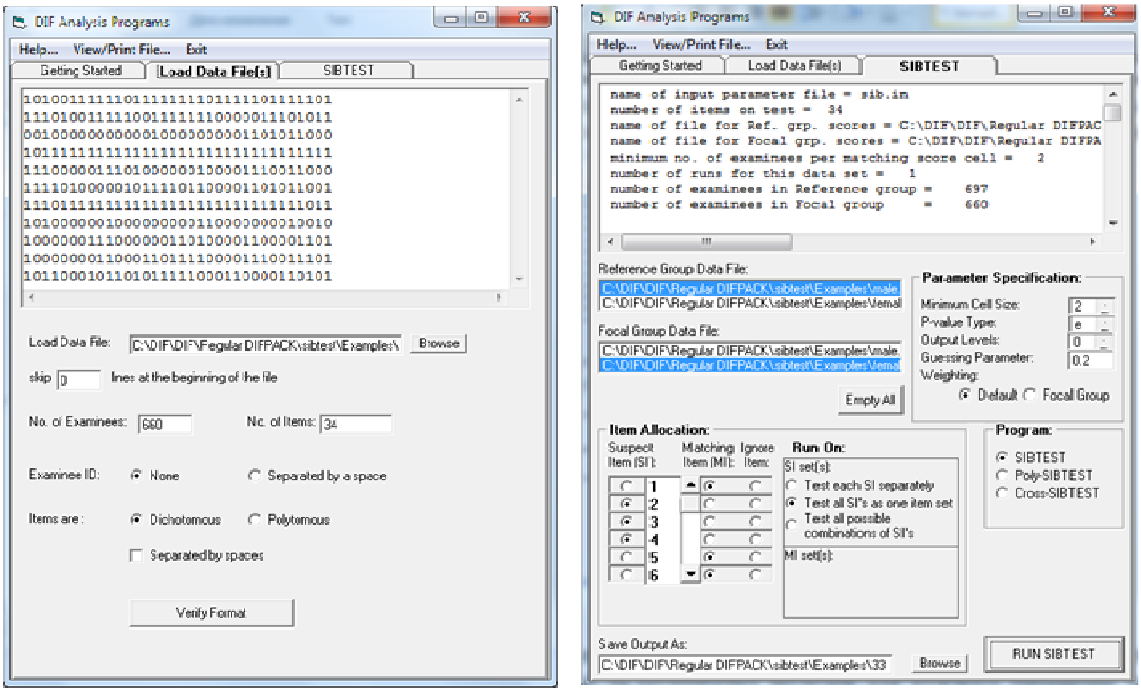


Рис. 1. Робоче вікно програми Easy-DIF



а) *б)*

Рис. 2. Робочі вікна програм пакету DIFPAK

Метод SIBTEST і його модифікації реалізовано у програмах пакету DIFPAK v.1.7 (Stout W., 2005). Пакет має три модулі: SIBTEST – для виявлення рівномірного DIF у тестах з дихотомічними завданнями; Poly-SIBTEST – для дослідження політомічних завдань, а також тестів, які містять завдання обох типів; Crossing-SIBTEST – для виявлення нерівномірного DIF у дихотомічних завданнях. Усі вони мають можливості для вибору параметрів аналізу і різних налаштувань за бажанням дослідника. В одному з вікон (рис. 2, а) потрібно завантажити окремо файли для референтної і фокусної груп і здійснити верифікацію даних. У наступному вікні (рис. 2, б) обираються всі опції для аналізу, а після запуску програми тут же можна бачити результат у вигляді звичного набору статистик, які обчислює даний метод, похибок і рівнів значущості. Завантажити пакет можна за посиланням http://psychometrictools.measuredprogress.org/dif1.

Програма IRTLRDIF v.2.0 (Thissen D., 2001) реалізує порівняння спрощеної і розширеної моделей для кожного параметра за методом IRT-LR. Вона допускає використання моделей Бірнбаума для завдань множинного вибору і Graded Model (Samejima F., 1969) для політомічних завдань. На рис. 3 представлено алгоритм попарного порівняння моделей для дихотомічних завдань, який реалізовано в даній програмі.

Спочатку порівнюються моделі A, у якій всі параметри вважаються однаковими в обох групах, і B, у якій усі параметри деякого вибраного завдання оцінюються окремо в різних групах (вільні). Якщо відповідна величина G^2 не перевищує 3.84 (найменше критичне значення розподілу χ^2 з одним ступенем волі при $\alpha = 0.05$), то таке завдання функціонує однаково в обох групах, жоден з параметрів не може бути причиною DIF. Якщо $G^2 > 3.84$, то порівнюються моделі далі, щоб встановити, який із параметрів є причиною DIF. Для виявлення DIF через параметр угадування *c* порівнюють моделі C і B. У моделі C параметр угадування вважається однаковим у обох групах, вона є спрощеною відносно B. Аналогічно порівнюють моделі D і C для виявлення нерівномірного DIF, що виникає через параметр роздільної здатності *a* за умови, що параметр *c* однаковий у обох групах. Порівнюючи моделі A і D, виясняють, чи не буде причиною DIF параметр складності *b* за умови, що параметри *a* і *c* залишаються однаковими у групах. Для двопараметричної моделі у дослідженні *a*-DIF порівнюють моделі E і B, а в дослідженні *b*-DIF за умови, що параметр *a* незмінний, порівнюють моделі A і E (рис. 3). У програмі можна створювати множину якірних завдань, у яких відсутність DIF є очевидною, а інші з ними порівнювати. Це підвищує якість аналізу.

IRTLRDIF створювалась для роботи в DOS, але добре працює в режимі командного рядка Windows. Вихідний текстовий документ добре пристосований для перенесення даних у різні статистичні пакети для подальшої роботи. Крім значень відповідних статистик і рівнів значущості, у вихідному документі виводяться параметри моделі для всіх завдань у фокусній і референтній групах з кожного кроку порівняння, що дозволяє будувати й аналізувати характеристичні криві. Завантажити програму можна зі сторінки автора http://www.unc.edu/~dthissen/dl.html.

Програм для дослідження DIF за методом LR розроблялось менше, оскільки його можна реалізувати самостійно у статистичних пакетах типу SPSS, SAS або R. Наприклад, у SPSS для цього використовуємо пункт меню `Анализ→Регрессия→ Логистическая…` і послідовно будуємо три логістичні регресії, вибираючи як коваріат спочатку загальний бал, далі змінну, що вказує на членство в групі, потім їх добуток. Обчислюємо різниці між значеннями `Chi-square` для кожної моделі. Якщо в порівнянні моделей 2 і 1 (3 і 2) ця різниця потрапляє в критичну область розподілу χ^2 з одним ступенем волі, маємо рівномірне (нерівномірне) DIF. SPSS не обчислює Weighted-Least-Squares R^2 для визначення розміру DIF, тому можна використати

Nagelkerke R^2. Щоб не повторювати цю процедуру для кожного завдання, можна скористатись макросами, розробленими Zumbo B. (1999) для дихотомічних іполітомічних завдань: http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html. Останнім часом з'явилося кілька реалізацій методу логістичної регресії в R, наприклад, пакет «lordif» (Choi S., Gibbons L., Crane P., 2011).

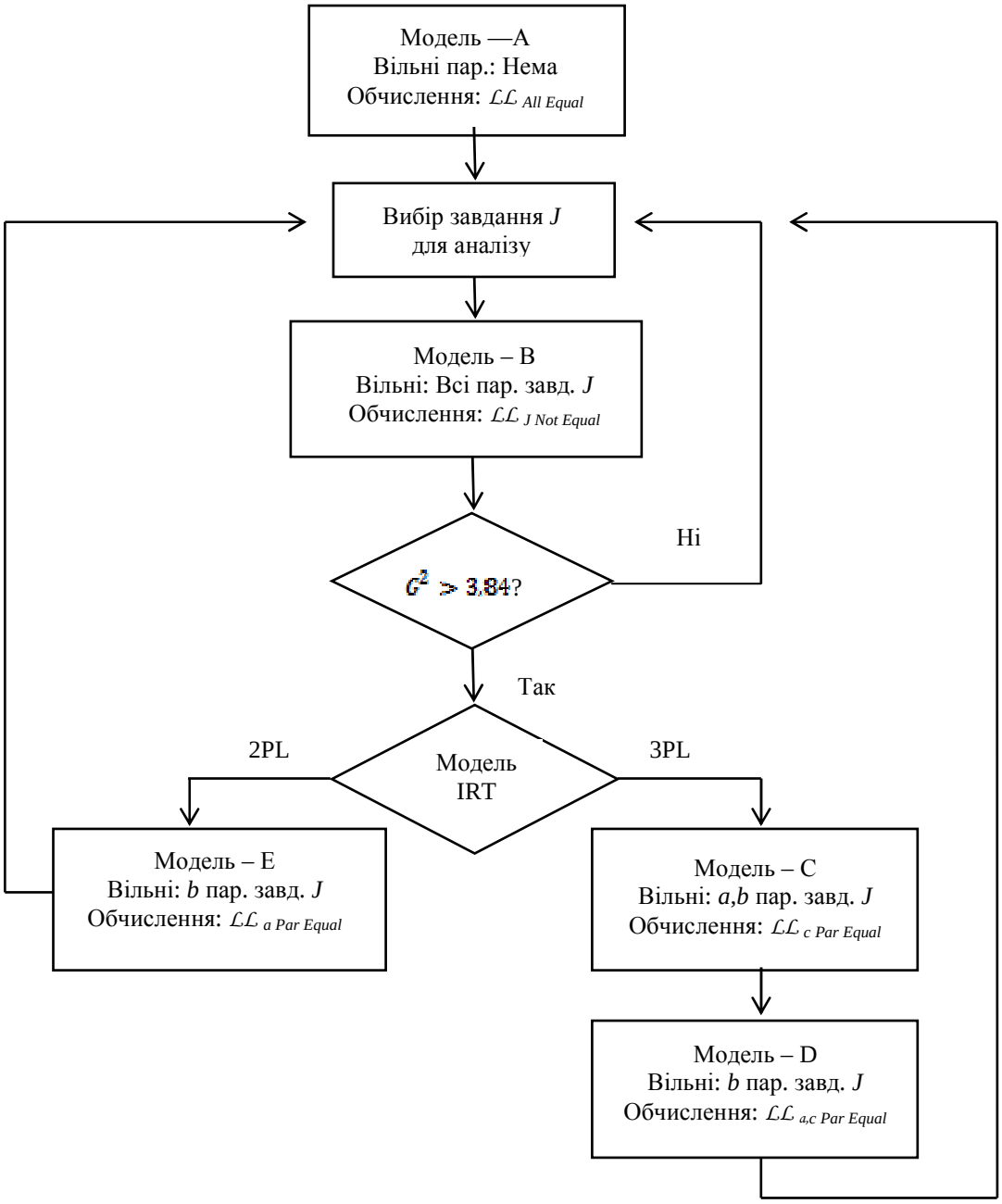


Рис. 3. Алгоритм послідовного порівняння розширеної і спрощеної моделей у програмі IRTLRDIF

Усі розглянуті засоби можуть успішно використовуватись для аналізу реальних даних, оскільки вони практично не мають обмежень на кількість учасників, але через свою доступність деякі з них є дуже корисними й незамінними для начальних цілей.

Для порівняння вищезгаданих методів використовувались реальні дані TIMSS з математики. Досліджувались 32 завдання, які виконували учасники трьох країн: України (239 учасників), Росії (349) та Сполучених Штатів (742). Кількість учасників з України є мінімальна з допустимих для деяких методів. Наприклад, MH вимагає мінімум 200 учасників у кожній групі, LR – 250. Тому результати за різними методами можуть суттєво різнитися. Попри це, на висновки може впливати співвідношення між розмірами вибірок, яке у нашому випадку не є критичним. Для кожної пари порівнянь було використано всі чотири методи. Аналіз підтвердив деякі очікувані й відомі факти.

Методом IRT-LR в усіх порівняннях була виявлена найбільша кількість завдань з DIF хоча б якого розміру (на рівні значущості 0.05 відхилено нульову гіпотезу про його відсутність): від 47% у парі Україна – Росія до 75% у парі Росія – США. Разом з тим, завдань зі значним DIF рівня C у кожній парі даний метод виявив набагато менше, ніж непараметричні методи MH і SIBTEST. Метод SIBTEST виявився найбільш чутливим у визначенні DIF рівня C: від 25% у парі Україна – Росія до 47% у парі Росія – США. Така чутливість даного методу також зазначена іншими авторами [8, 19]. Натомість, метод LR був найменш чутливим і при відхиленні нульової гіпотези, і при виявленні значного DIF рівня C: від 3% у парі Україна – Росія до 16% у парі Росія – США. У кожній парі порівнянь були завдання (від 1 до 3), у яких було виявлено нерівномірне DIF лише одним методом LR. Утім, в усіх завданнях, де метод LR виявив рівномірне DIF, усі інші методи також виявили DIF хоча б якогось розміру. Якщо ж методом LR було виявлено значне DIF рівня C, то всі інші методи приводили до такого ж висновку.

Важливими показниками узгодженості рішень за різними методами є відсоток збігів і кореляція між розмірами DIF. Відсоток збігів для двох методів вказує на частку завдань, у яких DIF виявлено обома методами, серед усіх завдань, у яких виявлено DIF. Відсоток збігів за всіма методами був вищий, чим більшого розміру вибірки порівнювались. Найбільш узгоджені результати демонстрували методи MH і SIBTEST (до 90% у парі Росія – США). Найменший відсоток збігів спостерігався між методами MH і LR (59% у парі Україна – США). Коефіцієнт кореляції між розмірами DIF для всіх пар методів у всіх парах порівнянь був вищим за 0.7. Такі показники узгодженості є задовільними, але далеко не ідеальними. Отже, з метою уникнення помилкових висновків варто використовувати кілька методів одночасно.

4. ВИСНОВКИ

Питання виявлення упередженого функціонування завдань і тесту (DIFi DTF) є важливими з точки зору забезпечення справедливого оцінювання відносно всіх груп учасників тестування. Потужний арсенал математичних і статистичних методів для виявлення DIF допомагає розробникам і користувачам тестів отримувати справедливі оцінки учасників, а сам тест вдосконалювати до перетворення його на об’єктивний інструмент вимірювання.

Неможливо уявити застосування розглянутих у роботі статистичних методів без спеціального програмного забезпечення. Не всі відомі статистичні пакети дозволяють проводити такий специфічний комплексний аналіз, тому було розроблено і продовжують розроблятись спеціалізовані пакети, серед яких є немало у вільному доступі для всіх бажаючих. Деякі з них цінні з методичної точки зору у вивченні студентами методів виявлення DIF в умовах обмеженого часу і ресурсів.

Проведений аналіз показав, що для підвищення надійності статистичних висновків доцільно одночасно застосовувати кілька різних методів. Особливо корисно поєднувати в одному дослідженні параметричні і непараметричні методи, якщо великі за обсягами вибірки є достатніми для цього. Але варто розуміти, що самі лише

статистичні висновки про наявність DIF ще не є вироком для такого завдання. Тільки після ґрунтовного аналізу статистичних результатів, отриманих різними методами і засобами, експертами з предметної галузі й фахівцями з тестування може прийматись рішення про те, що завдання дійсно призводить до *bias*. Експерти повинні чітко усвідомлювати причини, що призвели до цього, щоб вказати можливі шляхи вирішення конфліктної ситуації.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Лісова Т. В. DIF та DPF аналіз тесту на креативну самодостатність за допомогою програми Winsteps / Тетяна Лісова // Гуманітарний вісник. Тематичний випуск «Міжнародні Челпанівські психолого-педагогічні читання». – 2013. – Додаток 1 до Вип. 29, Том 1. – С. 255–261.
2. Фройнд Ф. А. Анализ DIF в оценке общего интеллекта для генерируемых компьютером графических тестовых задач в двух этнически различных выборках / [Ф. А. Фройнд, С. В. Давыдов, Й. П. Бертлинг, Х. Холлинг, Г. С. Шляхтин] // Социология. Психология. Философия. Вестник Нижегородского университета им. Н.И. Лобачевского. – 2012. – № 5 (1). – С. 334–341.
3. Acar T. Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR / T. Acar, H. Kelecioglu // Educational Sciences: Theory & Practice. – 2010. – 10 (2). – P. 639–649.
4. Ayala R. J. The Theory and Practice of Item Response Theory / R. J. de Ayala. – New York: Guilford Publications Incorporated, 2009. – 448 p.
5. Bolt D. M. A monte carlo comparison of parametric and nonparametric polytomous DIF detection methods / D. M. Bolt // Applied Measurement in Education. – 2002. – Vol. 15. – P. 113–141.
6. Camilli G. Test fairness / Gregory Camilli / In R. Brennan (Ed.), Educational measurement. – Westport, CT: ACE, Praeger series on higher education, 2006. – P. 221–256.
7. Dorans N.J. ETS Contributions to the Quantitative Assessment of Item, Test, and Score Fairness / N.J. Dorans. – Educational Testing Service: Princeton, New Jersey, 2013. – 38 p.
8. Ercikan K. Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests / [K. Ercikan, M. J. Gierl, T. McCreith et al.[// Applied Measurement in Education. – 2004. – Vol. 17(3). – P. 301–321.
9. Ercikan K. Examining the Construct Comparability of the English and French Versions of TIMSS / K. Ercikan, K. Koh // International Journal of Testing. – 2005. – Vol. 5(1). – P. 23–35.
10. Gierl M.J. Performance of SIBTEST When the Percentage of DIF Items is Large / M. J. Gierl, A. Gotzmann, K. A. Boughton // Applied Measurement in Education. – 2004. – Vol. 17(3). – P. 241–264.
11. González A. EASY-DIF: Software for Analyzing Differential Item Functioning Using the Mantel-Haenszel and Standardization Procedures / [A. González, J. L. Padilla, M. D. Hidalgo et al.[// Applied Psychological Measurement. – 2011. – Vol. 35(6). – P. 483–484.
12. Hambleton R. K. Translating achievement tests for use in cross-cultural studies / R.K. Hambleton // European Journal of Psychological Assessment. – 1993. – Vol. 9. – P. 57–68.
13. Holland P. W. Differential item performance and the Mantel-Haenszel procedure / P. W. Holland, D. T. In H. Wainer and H. Braun (ed), Test validity. – Hillsdale, NJ: Erlbaum, 1988. – P. 129–145.
14. Le L.T . Investigating Gender Differential Item Functioning Across Countries and Test Languages for PISA Science Items / Luc T. Le // International Journal of Testing. – 2009. – Vol. 9(2). – P. 122–133.
15. Oliveri M. E. Effects of Population Heterogeneity on Accuracy of DIF Detection / M. E. Oliveri, K. Ercikan, B. D. Zumbo // Applied Measurement in Education. – 2014. – Vol. 27(4). – P. 286–300.
16. Sandilands D. Investigating Sources of Differential Item Functioning in International Large-Scale Assessments Using a Confirmatory Approach / [D. Sandilands, M. E. Oliveri, B. D. Zumbo et al.] // International Journal of Testing. – 2013. – Vol. 13. – P. 152–174.
17. Shealy R. A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF / R. Shealy, W.F. Stout // Psychometrika. – 1993. – Vol.58. – P. 159–194.
18. SIBTEST manual [Електронний ресурс] / W. F. Stout, L. Roussos, 1995. – Режим доступу : http://psychometrictools.measuredprogress.org/dif1.
19. Stoneberg B. D. A study of gender-based and ethnic-based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement tests applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel chi-square Test [Електронний ресурс]. – Режим доступу : http://files.eric.ed.gov/fulltext/ED489949.pdf.
20. Swaminathan H. Detecting differential item functioning using logistic regression procedures / H. Swaminathan, H. J. Rogers // Journal of Educational Measurement. – 1990. – Vol. 27. – P. 361–370.

21. Thissen D. Detection of differential item functioning using the parameters of item response models / D. Thissen, L. Steinberg, H. Wainer // In P.W. Holland and H. Wainer (ed), Differential item functioning. – Hillsdale, NJ: Erlbaum, 1993. – P. 67–113.
22. Thissen, D. IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [Електронний ресурс]. – Режим доступу : http://www.unc.edu/~dthissen.
23. Zumbo B. D. A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores [Електронний ресурс]. – Режим доступу : http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html.
24. Zumbo B. D. Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going / B. D. Zumbo // Language Assessment Quarterly. – 2007. – Vol. 4(2). – P. 223–233.

Матеріал надійшов до редакції 29.09.2015 р.

МЕТОДЫ И ПРОГРАММНЫЕ СРЕДСТВА ИССЛЕДОВАНИЯ ВАЛИДНОСТИ ТЕСТОВЫХ РЕЗУЛЬТАТОВ ДЛЯ ГРУПП ТЕСТИРУЕМЫХ С ОПРЕДЕЛЕННЫМИ ИНДИВИДУАЛЬНЫМИ ОСОБЕННОСТЯМИ

Лисовая Татьяна Владимировна
доцент, кандидат физико-математических наук, доцент кафедры прикладной математики, информатики и образовательных измерений
Нежинский государственный университет имени Николая Гоголя, г. Нежин, Украина
tan-lisova@ndu.edu.ua

Аннотация. Необходимым условием наличия предвзятого оценивания некоторым тестом является дифференцированное (различное) функционирование заданий этого теста в отношении различных групп участников тестирования. В данной статье изложены идеи некоторых статистических методов выявления дифференцированного функционирования заданий, разработанных в рамках основных подходов к моделированию результатов тестирования: с помощью таблиц сопряженности, регрессионных моделей, многомерных моделей и моделей современной теории тестов IRT (Item Response Theory). Здесь рассматриваются методы Мантель-Ханзеля, логистической регрессии, SIBTEST и метод отношения правдоподобия. Указываются особенности и условия применения каждого метода. Проведен обзор существующих бесплатных программных средств, в которых реализованы данные методы. На примере реальных данных проведено сравнение этих методов. Указано на целесообразность использования нескольких методов одновременно с целью уменьшения риска ошибочных выводов.

Ключевые слова: дифференцированное функционирование задания; равномерное DIF; неравномерное DIF; логистическая регрессия; метод Мантель-Ханзеля; тест отношения правдоподобия.

METHODS AND SOFTWARE FOR RESEARCH OF THE VALIDITY OF TEST RESULTS FOR THE TESTED GROUPS WITH CERTAIN INDIVIDUAL CHARACTERISTICS

Tetiana V. Lisova
docent, PhD (physical and mathematical sciences), docent of the Department of Applied Mathematics, Computer Science and Educational Measurement
Nizhyn State Mykola Gogol University, Nizhyn, Ukraine
tan-lisova@ndu.edu.ua

Abstract. The necessary condition for the presence of biased assessment by some test is differential item functioning in different groups of test takers. The ideas of some statistical methods for detecting Differential Item Functioning are described in the given article. They were developed in the framework of the main approaches to modeling test results: using contingency

tables, regression models, multidimensional models and models of Item Response Theory. The Mantel-Haenszel procedure, logistic regression method, SIBTEST and Item Response Theory Likelihood Ratio Test are considered. The characteristics of each method and conditions of their application are specified. Overview of existing free software tools implementing these methods is carried out. Comparisons of these methods are conducted on the example of real data. Also notes that it is appropriate to use several methods simultaneously to reduce the risk of false conclusions.

Keywords: Differential Item Functioning; uniform DIF; non-uniform DIF; logistic regression; Mantel-Haenszel method; Likelihood Ratio Test.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Lisova T. V. DIF and DPF analysis of the test on creative self-sufficiency using Winsteps / Tetiana Lisova // Ghumanitarnyj visnyk. Special issue «International Chelpanov psychological and pedagogical reading». – 2013. – Annex 1 to the Issue 29, Vol. 1. – P. 255–261 (in Ukrainian).
2. Freund F. A. DIF analysis in the assessment of general intelligence for computer-generated graphics test tasks in two ethnically different samples / F. A. Freund, S. V. Davydov, J. P. Bertling, H. Holling, G. S. Shlyakhtin // Sociologija. Psihologija. Filosofija. Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo. – 2012. – Vol. 5 (1). – P. 334–341 (in Russian).
3. Acar T. Comparison of Differential Item Functioning Determination Techniques: HGLM, LR and IRT-LR / T. Acar, H. Kelecioglu // Educational Sciences: Theory & Practice. – 2010. – 10 (2). – P. 639–649 (in English).
4. Ayala R. J. The Theory and Practice of Item Response Theory / R. J. de Ayala. – New York: Guilford Publications Incorporated, 2009. – 448 p. (in English).
5. Bolt D. M. A monte carlo comparison of parametric and nonparametric polytomous DIF detection methods / D.M. Bolt // Applied Measurement in Education. – 2002. – Vol. 15. – P. 113–141 (in English).
6. Camilli G. Test fairness / Gregory Camilli / In R. Brennan (Ed.), Educational measurement. – Westport, CT: ACE, Praeger series on higher education, 2006. – P. 221–256 (in English).
7. Dorans N. J. ETS Contributions to the Quantitative Assessment of Item, Test, and Score Fairness / N. J. Dorans. – Educational Testing Service: Princeton, New Jersey, 2013. – 38 p. (in English).
8. Ercikan K. Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests / K. Ercikan, M. J. Gierl, T. McCreith et al. // Applied Measurement in Education. – 2004. – Vol. 17(3). – P. 301–321 (in English)..
9. Ercikan K. Examining the Construct Comparability of the English and French Versions of TIMSS / K. Ercikan, K. Koh // International Journal of Testing. – 2005. – Vol. 5(1). – P. 23–35 (in English).
10. Gierl M. J. Performance of SIBTEST When the Percentage of DIF Items is Large / M. J. Gierl, A. Gotzmann, K. A. Boughton // Applied Measurement in Education. – 2004. – Vol. 17(3). – P. 241–264 (in English).
11. González A. EASY-DIF: Software for Analyzing Differential Item Functioning Using the Mantel-Haenszel and Standardization Procedures / A. González, J. L. Padilla, M. D. Hidalgo et al. // Applied Psychological Measurement. – 2011. – Vol. 35(6). – P. 483–484 (in English).
12. Hambleton R. K. Translating achievement tests for use in cross-cultural studies / R. K. Hambleton // European Journal of Psychological Assessment. – 1993. – Vol. 9. – P. 57–68 (in English).
13. Holland P. W. Differential item performance and the Mantel-Haenszel procedure / P. W. Holland, D. T. Thayer / In H. Wainer and H. Braun (ed), Test validity. – Hillsdale, NJ: Erlbaum, 1988. – P. 129–145 (in English).
14. Le L.T. Investigating Gender Differential Item Functioning Across Countries and Test Languages for PISA Science Items / Luc T. Le // International Journal of Testing. – 2009. – Vol. 9(2). – P. 122–133. (in English)
15. Oliveri M. E. Effects of Population Heterogeneity on Accuracy of DIF Detection / M. E. Oliveri, K. Ercikan, B. D. Zumbo // Applied Measurement in Education. – 2014. – Vol. 27(4). – P. 286–300 (in English).
16. Sandilands D. Investigating Sources of Differential Item Functioning in International Large-Scale Assessments Using a Confirmatory Approach / D. Sandilands, M. E. Oliveri, B. D. Zumbo et al. // International Journal of Testing. – 2013. – Vol. 13. – P. 152–174 (in English).
17. Shealy R. A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF / R. Shealy, W. F. Stout // Psychometrika. – 1993. – Vol.58. – P. 159–194 (in English).
18. SIBTEST manual [online] / W.F. Stout, L. Roussos, 1995. – Available from : http://psychometrictools.measuredprogress.org/dif1 (in English).

19. Stoneberg B. D. A study of gender-based and ethnic-based Differential Item Functioning (DIF) in the Spring 2003 Idaho Standards Achievement tests applying the Simultaneous Bias Test (SIBTEST) and the Mantel-Haenszel chi-square Test [online]. – Available from : http://files.eric.ed.gov/fulltext/ED489949.pdf (in English).
20. Swaminathan H. Detecting differential item functioning using logistic regression procedures / H. Swaminathan, H. J. Rogers // Journal of Educational Measurement. – 1990. – Vol. 27. – P. 361–370 (in English).
21. Thissen D. Detection of differential item functioning using the parameters of item response models / D. Thissen, L. Steinberg, H. Wainer // In P. W. Holland and H. Wainer (ed), Differential item functioning. – Hillsdale, NJ: Erlbaum, 1993. – P. 67–113 (in English).
22. Thissen, D. IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning [online]. – Available from : http://www.unc.edu/~dthissen (in English)
23. Zumbo B. D. A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores [online]. – Available from : http://www.educ.ubc.ca/faculty/zumbo/DIF/index.html (in English).
24. Zumbo B. D. Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going / B. D. Zumbo // Language Assessment Quarterly. – 2007. – Vol. 4(2). – P. 223–233 (in English).

(cc) BY-NC-SA

This work is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.