

Ж.Ж. Бейсенова^{*1}, Л.Е. Шинетова², Г.К. Жаббаева³, С.Ж. Анафияева⁴
^{1,2,3,4}РГП на ПХВ «Национальный центр тестирования» МНВО РК, г. Астана, Республика Казахстан
^{*}e-mail: zhanar-beisenova@mail.ru

¹ORCID 0000-0002-7347-1538, ²ORCID 0000-0003-4280-7999,

³ORCID 0009-0006-1003-6316, ⁴ORCID 0009-0000-9124-9976

ДИФФЕРЕНЦИАЛЬНОЕ ФУНКЦИОНИРОВАНИЕ ТЕСТОВЫХ ЗАДАНИЙ В ТЕСТИРОВАНИИ ПЕДАГОГОВ НА ОСНОВЕ МОДЕЛИ РАША

В статье представлены результаты психометрического анализа дифференциального функционирования заданий (DIF) по признаку пола на материале трех параллельных вариантов теста для педагогов. Цель исследования - оценить наличие, величину и направление DIF-эффектов в рамках модели Раша и определить их статистическую устойчивость при контроле множественных сравнений. Предполагалось, что при сопоставимой трудности параллельных вариантов доля и направление DIF-эффектов будут ограничены отдельными заданиями и не образуют системного смещения в пользу одного пола. Материалом послужили результаты тестирования 1417 испытуемых (1071 женщина, 346 мужчин), выполнявших три варианта теста по 50 дихотомически оцениваемых заданий каждый; обработка данных проведена в программе Winsteps 5.8.3.0. Дифференциальное функционирование оценивалось по величине DIF, с дополнительной проверкой по статистике Mantel-Haenszel и контролем ложных обнаружений (FDR) по процедуре Бенджамини-Хохберга, примененной отдельно в каждом варианте. По первичному критерию DIF выявлен в 25 из 150 позиций «задание × вариант» (16,7%); после FDR-коррекции статистически устойчивыми остались 9 позиций (6,0%). Направление устойчивых эффектов оказалось практически сбалансированным: 4 задания труднее для женщин, 5 - для мужчин, что не дает оснований говорить о системном смещении трудности в пользу одного пола. Наибольшая концентрация DIF-сигналов отмечена в варианте 703. Результаты в целом подтверждают приемлемый уровень измерительной инвариантности использованных вариантов теста по признаку пола и позволяют сформулировать рекомендации по регулярному мониторингу DIF при разработке и эксплуатации параллельных форм теста.

Ключевые слова: дифференциальное функционирование заданий (DIF), модель Раша, инвариантность измерения, тестирование педагогов, гендерные различия, статистика Mantel-Haenszel.

Введение

Справедливость и сопоставимость результатов стандартизированного тестирования педагогов напрямую зависят от того, насколько тестовые задания одинаково функционируют для различных подгрупп испытуемых при одном и том же уровне измеряемой способности. Это свойство в литературе по психометрическому анализу принято называть инвариантностью измерения, а его нарушение на уровне отдельного задания - дифференциальным функционированием задания, или DIF (Differential Item Functioning) [1, 2]. Модель Раша, лежащая в основе большинства современных систем педагогических измерений, включая аналитическую платформу Winsteps, задает естественную основу для проверки инвариантности: если параметр трудности задания статистически различается для двух групп испытуемых, сопоставимых по общему уровню способности, это рассматривается как индикатор возможного DIF [1, 3].

Начиная с работ Holland и Thayer, статистика Mantel-Haenszel стала одним из наиболее распространенных непараметрических инструментов обнаружения DIF наряду с методами, основанными на модели Раша [4, 5]. Накопленный за последние десятилетия методический опыт показывает, что в задачах с большим числом одновременных сравнений использование только номинального уровня значимости $p < 0,05$ приводит к завышению доли ложных

срабатываний; для контроля этого эффекта рекомендуется процедура контроля ложных обнаружений (False Discovery Rate, FDR), предложенная Бенджамини и Хохбергом [7]. Современные обзоры методов DIF подтверждают, что сочетание содержательного критерия величины эффекта (например, порогового значения DIF) со статистическим контролем множественных сравнений повышает надежность выводов и снижает риск избыточной интерпретации случайных колебаний как содержательного смещения [7, 8].

Гендерный DIF остается одной из наиболее изучаемых тем прикладной психометрики. Исследования на материале крупномасштабных оценочных программ - от PISA до национальных экзаменов - показывают, что гендерные различия в выполнении отдельных заданий существуют, но носят преимущественно локальный, а не системный характер и заметно зависят от предметной области, формата задания и особенностей выборки [9-13]. Эти результаты подчеркивают необходимость эмпирической проверки DIF для каждого конкретного инструмента, а не экстраполяции выводов, полученных на другом материале.

В казахстанском контексте вопросы психометрического качества и справедливости инструментов оценивания приобретают дополнительную практическую значимость в связи с реформированием системы оценки знаний педагогов и обучающихся [14, 15], а нормативную основу для разработки и применения стандартизированных инструментов оценивания в Республике Казахстан задают соответствующие акты уполномоченного органа в сфере образования [16]. Вместе с тем систематические публикации, в которых DIF-анализ проводился бы одновременно на нескольких параллельных вариантах одного теста для педагогов с явным контролем множественных сравнений, в отечественной литературе представлены ограниченно, что определяет актуальность настоящего исследования.

Объект исследования - результаты выполнения тестовых заданий испытуемыми, различающимися по признаку пола. Предмет исследования - дифференциальное функционирование тестовых заданий (DIF) по признаку пола в трех параллельных вариантах теста для педагогов, рассматриваемое в рамках модели Раша.

Цель исследования - оценить наличие, величину, направление и статистическую устойчивость дифференциального функционирования заданий по признаку пола в трех параллельных вариантах теста для педагогов. Для достижения цели решались следующие задачи: 1) оценить достаточность и структуру выборки по вариантам теста; 2) выявить задания с признаками DIF по критерию DIF Contrast и статистике Rasch-Welch с сопоставлением со статистикой Mantel-Haenszel; 3) применить процедуру контроля ложных обнаружений (FDR) для повышения устойчивости выводов; 4) описать направление выявленных эффектов и сопоставить его с данными других исследований гендерного DIF.

Научная новизна работы состоит в применении комплексного подхода - сочетания критерия величины эффекта, статистики Rasch-Welch, статистики Mantel-Haenszel и FDR-коррекции - к трем параллельным вариантам теста оценки знаний педагогов, используемого в системе оценивания Республики Казахстан. Практическая значимость заключается в том, что результаты могут быть использованы для приоритизации содержательной экспертизы отдельных заданий и совершенствования процедур контроля качества параллельных форм теста.

Материалы и методы исследования

Материалом исследования послужили результаты тестирования, выполненного по трем параллельным вариантам теста для педагогов (условные номера вариантов: 671, 703 и 078), каждый из которых включает 50 дихотомически оцениваемых заданий закрытого типа. Общий объем выборки составил 1417 наблюдений, из них 1071 - женщины (75,6%) и 346 - мужчины (24,4%). Кодировка группы пола в аналитических таблицах Winsteps: D - женский пол, M - мужской пол.

Численность и состав выборки, а также показатели надежности использованных вариантов теста приведены в таблице 1.

Таблица 1 - Характеристика выборки и показатели надежности по трем вариантам теста

Вариант	N	Женщины (D), n	Мужчины (M), n	Женщины, %	Мужчины, %	KR-20	Person reliability
671	302	199	103	65,9	34,1	0,88	0,87
703	488	368	120	75,4	24,6	0,92	0,89
078	627	504	123	80,4	19,6	0,90	0,88
Итого	1417	1071	346	75,6	24,4	0,80-0,92	0,87-0,89
Примечание - составлено авторами на основе данных Winsteps, Table 3.1							

Объем выборок по вариантам составляет от 302 до 627 человек, что в целом соответствует объемам, рекомендуемым для устойчивого DIF-анализа непараметрическими и Rasch-based методами [7]. Вместе с тем численность мужской подгруппы (от 103 до 123 человек по вариантам) заметно уступает численности женской подгруппы, что закономерно отражает гендерный состав контингента педагогов и является типичной особенностью выборок в исследованиях этой профессиональной группы [11]. Такая асимметрия увеличивает стандартную ошибку DIF-оценок для менее многочисленной группы и требует осторожности при интерпретации эффектов, близких к пороговому значению. Об этом подробнее нами будет рассмотрено в разделе «Заключение».

Расчеты выполнены в программе Winsteps 5.8.3.0 в рамках однопараметрической модели Раша [1, 2]. Основным источником данных для DIF-анализа послужила Table 30.1 Winsteps, содержащая для каждого задания групповые оценки трудности (DIF Measure) для женщин (D) и мужчин (M), величину DIF Contrast, ее стандартную ошибку, значение статистики Rasch-Welch t и соответствующий p -value, а также статистику Mantel-Haenszel. Table 3.1 использовалась для оценки объема выборки и показателей надежности теста (KR-20, Person reliability).

Поскольку в выводимых таблицах Winsteps первой указывается группа D, величина DIF Contrast рассчитывалась как разность D - M. Положительное значение DIF Contrast означает, что задание оценено как более трудное для женщин и относительно благоприятнее для мужчин; отрицательное значение означает обратное соотношение.

В качестве первичного критерия выявления DIF использовалось сочетание содержательного и статистического порогов: абсолютная величина $|DIF Contrast| \geq 0,50$ логита при $p < 0,05$ по статистике Rasch-Welch. Пороговое значение 0,50 логита соответствует ориентиру, рекомендуемому руководством Winsteps для заметного DIF-эффекта, и характеризует эффект как минимум умеренной величины [2, 9]. Дополнительно для каждого задания рассматривалось согласие с результатом статистики Mantel-Haenszel [4, 5], что позволяет оценить устойчивость сигнала при использовании методически независимого подхода.

Поскольку в каждом варианте одновременно проверялось 50 заданий, для контроля вероятности ложных обнаружений применялась процедура Бенджамини-Хохберга (FDR) при уровне $q < 0,05$ [6]. FDR рассчитывался отдельно внутри каждого варианта по p -value статистики Rasch-Welch. Такой раздельный расчет учитывает, что варианты представляют собой независимые наборы гипотез.

Наличие статистически значимого и устойчивого после FDR-коррекции DIF интерпретируется в настоящей работе как свидетельство статистической неинвариантности параметра трудности задания между женщинами и мужчинами при контроле общего уровня измеряемой способности. Такой результат не интерпретируется как доказанная содержательная или гендерная предвзятость конкретного задания, поскольку текст задания,

его тематическая привязка и когнитивная структура в рамках настоящего анализа не рассматривались; это ограничение согласуется с общепринятой методической позицией в литературе по DIF [7, 8].

Результаты и обсуждение

Сводные результаты DIF-анализа по трем вариантам теста представлены в таблице 2.

Таблица 2 - Сводные результаты DIF-анализа по трем вариантам теста

Вариант	Заданий	DIF по первичному критерию	Доля, %	Сохранились после FDR
671	50	5	10,0	1
703	50	13	26,0	5
078	50	7	14,0	3
Итого	150	25	16,7	9 (6,0%)
Примечание - составлено авторами на основе данных Winsteps, Table 30.1				

По первичному критерию DIF выявлен в 25 из 150 позиций «задание × вариант» (16,7%). Из них 17 позиций одновременно имели $p < 0,05$ по статистике Mantel-Haenszel, что говорит о согласованности результатов двух методически различных подходов для большинства выявленных сигналов. После контроля FDR устойчивыми остались 9 позиций (6,0% от общего числа проверенных заданий). Наибольшая концентрация DIF-сигналов как по первичному критерию, так и после FDR-коррекции наблюдалась в варианте 703 (см. рисунок 1).

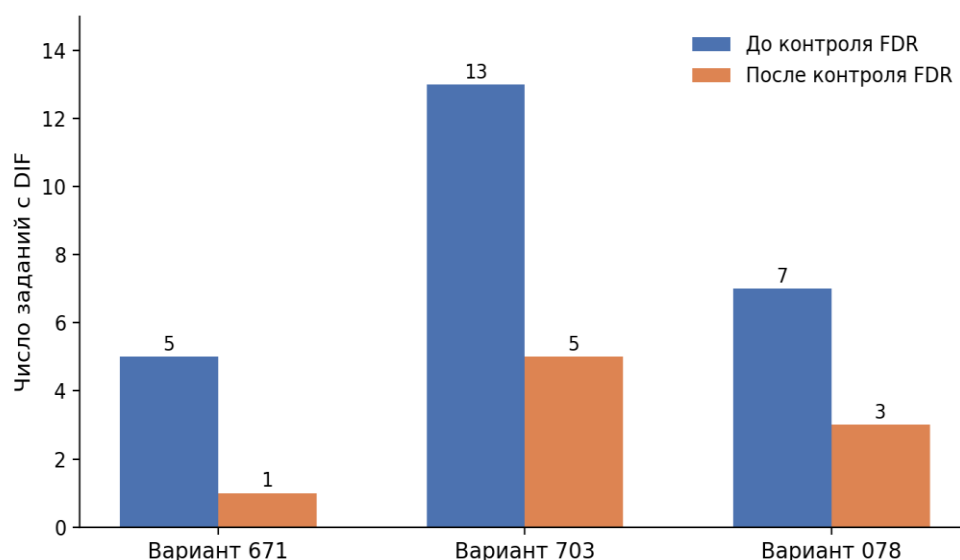


Рисунок 1 - Число заданий с DIF по вариантам теста, до и после контроля FDR

Примечание - составлено авторами на основе данных таблицы 2

Направление DIF-эффектов по вариантам приведено в таблице 3.

Таблица 3 - Направление DIF-эффектов по признаку пола

Вариант	DIF всего	Труднее для женщин (+)	Труднее для мужчин (-)	После FDR
671	5	2	3	1 (0+ / 1-)
703	13	7	6	5 (3+ / 2-)
078	7	4	3	3 (1+ / 2-)
Итого	25	13	12	9 (4+ / 5-)

Примечание - составлено авторами на основе данных таблицы 30.1 Winsteps

При первичном скрининге направление DIF в совокупности трех вариантов оказалось практически сбалансированным: 13 эффектов положительные (задание труднее для женщин) и 12 отрицательных (задание труднее для мужчин). Среди 9 эффектов, сохранившихся после FDR-коррекции, 4 положительные и 5 отрицательных. Распределение отдельных значений DIF Contrast относительно порогового уровня $\pm 0,50$ логита показано на рисунке 2.

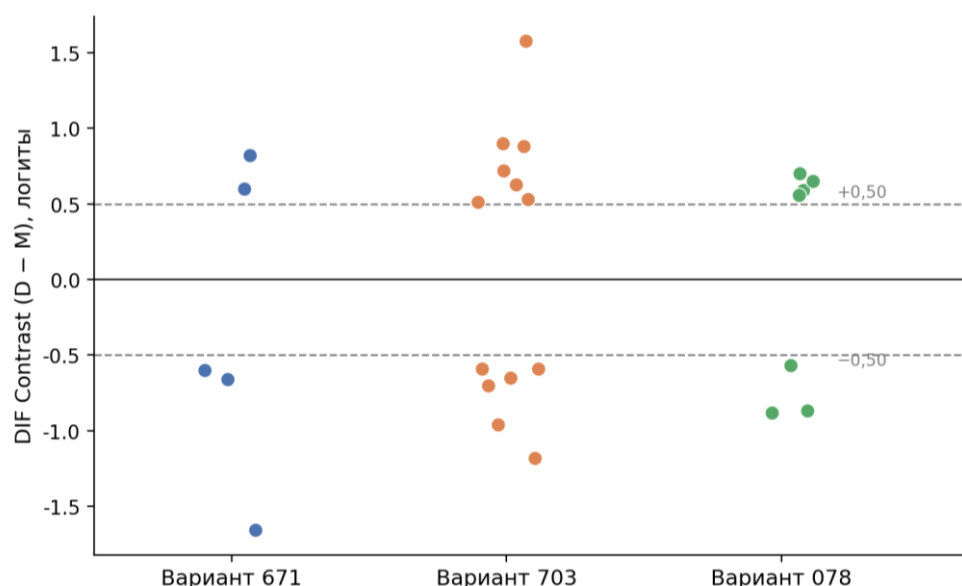


Рисунок 2 - Распределение DIF Contrast по трем вариантам теста
 (пунктирные линии - пороговый уровень $\pm 0,50$ логита; положительные значения соответствуют большей трудности для женщин, отрицательные - для мужчин)
 Примечание - составлено авторами на основе данных таблицы 30.1 Winsteps

Вариант 671.

По первичному критерию в варианте 671 выявлено 5 заданий с DIF (10,0% от общего числа): № 3, 11, 14, 20 и 34. Два из пяти сигналов также статистически значимы по критерию Mantel-Haenszel. После FDR-коррекции сохраняется одно задание - № 14 (таблица 4).

Таблица 4 - Задания с DIF в варианте 671

Задание	DIF женщин (D)	DIF мужчин (M)	DIF Contrast (D-M)	SE	p (Rasch-Welch)	q (FDR)
3	1,32	0,49	0,82	0,29	0,0044	0,1100
11	-0,99	-0,33	-0,66	0,28	0,0169	0,2817
14	1,68	3,34	-1,66	0,46	0,0004	0,0200
20	0,62	0,01	0,60	0,27	0,0276	0,3394
34	0,62	1,21	-0,60	0,29	0,0384	0,3394

Примечание - составлено авторами на основе данных таблицы 30.1 Winsteps

Наибольшая по модулю величина DIF в этом варианте зафиксирована для задания № 14 (DIF Contrast = -1,66 логита; $p = 0,0004$; сохраняется после FDR при $q = 0,0200$). Отрицательный знак указывает, что данное задание оценено как более трудное для мужчин, чем для женщин.

Вариант 703.

По первичному критерию в варианте 703 выявлено 13 заданий с DIF (26,0%): № 3, 9, 11, 12, 18, 22, 23, 27, 29, 39, 45, 47 и 49 - наибольшее число среди трех рассмотренных вариантов. Девять из тринадцати сигналов также статистически значимы по критерию Mantel-Haenszel. После FDR-коррекции сохраняются задания № 9, 11, 22, 23 и 39 (таблица 5).

Таблица 5 - Задания с DIF в варианте 703

Задание	DIF женщин (D)	DIF мужчин (M)	DIF Contrast (D-M)	SE	p (Rasch-Welch)	q (FDR)
3	-0,45	0,14	-0,59	0,26	0,0229	0,1145
9	-0,18	-1,06	0,88	0,28	0,0022	0,0220
11	4,07	2,49	1,58	0,32	<0,0001	<0,0001
12	-1,05	-0,46	-0,59	0,28	0,0349	0,1442
18	-0,34	-1,06	0,72	0,29	0,0123	0,0689
22	-0,51	0,45	-0,96	0,26	0,0002	0,0025
23	-0,09	1,10	-1,18	0,25	<0,0001	<0,0001
27	0,50	-0,13	0,63	0,25	0,0124	0,0689
29	0,22	-0,32	0,53	0,25	0,0372	0,1442
39	1,43	0,53	0,90	0,24	0,0002	0,0025
45	-0,97	-0,27	-0,70	0,27	0,0106	0,0689
47	0,27	0,92	-0,65	0,25	0,0086	0,0689
49	1,65	1,14	0,51	0,24	0,0375	0,1442

Примечание - составлено авторами на основе данных таблицы 30.1 Winsteps

Наибольшая по модулю величина DIF в варианте 703 зафиксирована для задания № 11 (DIF Contrast = +1,58 логита; $p < 0,0001$; сохраняется после FDR). Положительный знак указывает, что задание оценено как более трудное для женщин, чем для мужчин. Более

высокая частота DIF-сигналов в этом варианте по сравнению с двумя другими (26,0% против 10,0% и 14,0% по первичному критерию) позволяет рассматривать его как приоритетный объект для последующей содержательной экспертизы заданий, хотя сама по себе более высокая частота DIF не свидетельствует о предвзятости варианта в целом.

Вариант 078.

По первичному критерию в варианте 078 выявлено 7 заданий с DIF (14,0%): № 8, 15, 18, 27, 28, 29 и 34. Шесть из семи сигналов также статистически значимы по критерию Mantel-Haenszel. После FDR-коррекции сохраняются задания № 27, 29 и 34 (таблица 6).

Таблица 6 - Задания с DIF в варианте 078

Задание	DIF женщин (D)	DIF мужчин (M)	DIF Contrast (D-M)	SE	p (Rasch-Welch)	q (FDR)
8	1,39	0,80	0,59	0,23	0,0097	0,0970
15	-0,41	0,16	-0,57	0,24	0,0193	0,1457
18	2,37	1,81	0,56	0,24	0,0204	0,1457
27	-1,03	-0,16	-0,88	0,26	0,0008	0,0350
28	1,37	0,72	0,65	0,23	0,0044	0,0550
29	-1,22	-0,35	-0,87	0,27	0,0014	0,0350
34	0,95	0,25	0,70	0,23	0,0028	0,0467
Примечание - составлено авторами на основе данных таблицы 30.1 Winsteps						

Наибольшая по модулю величина DIF в варианте 078 зафиксирована для задания № 27 (DIF Contrast = -0,88 логита; $p = 0,0008$; сохраняется после FDR). Отрицательный знак указывает, что задание оценено как более трудное для мужчин, чем для женщин.

Обсуждение результатов

Полученная картина в целом согласуется с представлением о гендерном DIF, сложившимся в международной литературе: доля заданий с признаками DIF варьирует от теста к тесту и от варианта к варианту, эффекты чаще носят локальный, а не системный характер, а направление смещения не образует устойчивого перекоса в пользу одной из групп [9-11]. В настоящем исследовании после статистически более строгого критерия (контроль FDR) доля заданий с устойчивым DIF снизилась с 16,7% до 6,0%, что подчеркивает практическую значимость контроля множественных сравнений при анализе тестов с большим числом заданий [6, 7].

Различия между вариантами по частоте DIF - от 10,0% в варианте 671 до 26,0% в варианте 703 по первичному критерию - могут объясняться как особенностями конкретных заданий, так и структурой выборки (например, распределением уровня способности внутри гендерных подгрупп для данного варианта). Поскольку тематическое содержание заданий не анализировалось, в рамках настоящей работы невозможно установить содержательную причину более высокой частоты DIF-сигналов в варианте 703. Это направление обозначено как перспектива дальнейших исследований.

Сопоставление с работами, изучавшими источники гендерного DIF через анализ дистракторов и форматов заданий [10-12], указывает на то, что дальнейшее уточнение природы выявленных эффектов потребует привлечения содержательной экспертизы конкретных заданий, привязки к предметной области и, при необходимости, анализа типа

ответа. Такой шаг выходит за рамки настоящего психометрического анализа, но представляется логичным продолжением работы.

С практической точки зрения результаты имеют значение для процедур контроля качества параллельных форм теста в РГП на ПХВ «Национальный центр тестирования»: регулярный DIF-анализ по полу с обязательным контролем множественных сравнений позволяет заблаговременно выявлять задания, заслуживающие дополнительной содержательной экспертизы.

Заключение

Проведенный психометрический анализ трех параллельных вариантов теста для педагогов (671, 703, 078; $N = 1417$) в рамках модели Раша показал, что по первичному критерию ($|DIF\ Contrast| \geq 0,50$ логита, $p < 0,05$) дифференциальное функционирование по признаку пола выявляется в 25 из 150 позиций «задание $\times$ вариант» (16,7%), а после контроля ложных обнаружений (FDR) статистически устойчивым остается существенно меньшее число сигналов - 9 позиций (6,0%). Гипотеза о том, что выявленные эффекты будут носить локальный, а не системный характер, в целом подтвердилась: направление устойчивых после FDR эффектов оказалось практически сбалансированным (4 задания труднее для женщин, 5 - для мужчин), что не дает оснований говорить о системном смещении трудности использованных вариантов теста в пользу одного пола.

Полученные результаты имеют значение как для теории, так и для практики педагогических измерений. В теоретическом плане они подтверждают целесообразность сочетания содержательного порога величины эффекта, независимой проверки по статистике Mantel-Haenszel и статистического контроля множественных сравнений (FDR) при анализе DIF на тестах с большим числом одновременно проверяемых заданий. В практическом плане результаты позволяют рекомендовать: 1) приоритетную содержательную экспертизу заданий, сохранивших статистическую значимость DIF после FDR-коррекции, в первую очередь в варианте 703, где концентрация сигналов оказалась наибольшей; 2) включение DIF-анализа с контролем множественных сравнений в стандартную процедуру психометрической проверки параллельных форм теста; 3) регулярный мониторинг DIF по полу при последующих циклах тестирования педагогов для отслеживания устойчивости выявленных закономерностей во времени.

Ограничения исследования связаны с тем, что: содержание и тематическая привязка заданий не анализировались, в связи с чем выявленный DIF не может интерпретироваться как доказанная содержательная предвзятость конкретных заданий; численность мужской подгруппы заметно меньше численности женской подгруппы во всех трех вариантах, что увеличивает стандартные ошибки DIF-оценок и требует осторожной интерпретации эффектов, близких к пороговому значению; анализ ограничен тремя вариантами теста и не включает проверку устойчивости результатов во времени. Перспективным направлением дальнейших исследований представляется содержательный анализ заданий с устойчивым DIF, а также расширение выборки вариантов и циклов тестирования для повторной проверки выявленных закономерностей.

Информация о финансировании

Работа выполнена в рамках гранта внутреннего финансирования РГП на ПХВ «Национальный центр тестирования» МНВО РК «Психометрическое исследование качества тестовых заданий системы оценки знаний педагогов (ОЗП) и разработка рекомендаций по их совершенствованию».

Список литературы

1. Rasch G. Probabilistic Models for Some Intelligence and Attainment Tests. - Chicago : The University of Chicago Press, 1980 (переизд.). - 199 p.
2. Linacre J.M. Winsteps® Rasch Measurement Computer Program User's Guide. - Portland, Oregon: Winsteps.com, 2024. - <https://www.winsteps.com/winman/index.htm?differentialitemfunctioningdif.htm>
3. Hambleton R.K., Swaminathan H., Rogers H.J. Fundamentals of Item Response Theory. - Newbury Park, CA : Sage Publications, 1991. - 174 p.
4. Camilli G., Shepard L.A. Methods for Identifying Biased Test Items. - Thousand Oaks, CA : Sage Publications, 1994. - 174 p.
5. Holland P.W., Thayer D.T. Differential item performance and the Mantel-Haenszel procedure // Test Validity / eds. H. Wainer, H.I. Braun. - Hillsdale, NJ : Lawrence Erlbaum Associates, 1988. - P. 129-145.
6. Mantel N., Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease // Journal of the National Cancer Institute. - 1959. - Vol. 22, № 4. - P. 719-748.
7. Benjamini Y., Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing // Journal of the Royal Statistical Society: Series B (Methodological). - 1995. - Vol. 57, № 1. - P. 289-300. - DOI: 10.2307/2346101
8. Mapuranga R., Dorans N.J., Middleton K. A Review of Recent Developments in Differential Item Functioning : ETS Research Report RR-08-43. - Princeton, NJ : Educational Testing Service, 2008. - 32 p. - <https://doi.org/10.1002/j.2333-8504.2008.tb02129.x>
9. Monahan P.O., McHorney C.A., Stump T.E., Perkins A.J. Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression // Journal of Educational and Behavioral Statistics. - 2007. - Vol. 32, № 1. - P. 92-109. - <https://doi.org/10.3102/1076998606298035>
10. Liu Y., Yin H., Xin T., Shao L., Yuan L. A comparison of differential item functioning detection methods in cognitive diagnostic models // Frontiers in Psychology. - 2019. - Vol. 10. - Article 1137. - <https://doi.org/10.3389/fpsyg.2019.01137>
11. Liu O.L., Wilson M. Gender differences in large-scale math assessments: PISA trend 2000 and 2003 // Applied Measurement in Education. - 2009. - Vol. 22, № 2. - P. 164-184. - <https://doi.org/10.1080/08957340902754635>
12. Susongko P., Arfiani Y., Kusuma M. Determination of gender differential item functioning in Tegal-students' scientific literacy skills with integrated science (SLiSIS) test using Rasch model // Jurnal Pendidikan IPA Indonesia. - 2021. - Vol. 10, № 2. - P. 270-281. - <https://doi.org/10.15294/jpii.v10i2.26775>
13. Ozdemir B., AlGhamdi H.M. Investigating the distractors to explain DIF effects across gender in large-scale tests with non-linear logistic regression models // Frontiers in Education. - 2022. - Vol. 6. - Article 748884. - <https://doi.org/10.3389/educ.2021.748884>
14. Shear B.R. Gender bias in test item formats: Evidence from PISA 2009, 2012, and 2015 math and reading tests // Journal of Educational Measurement. - 2020. - Vol. 57, № 3. - P. 369-396. - <https://doi.org/10.1111/jedm.12372>
15. Карданова Е.Ю., Иванова А.Е. Психометрические исследования: современные методы и новые возможности для образования // Вопросы образования / Educational Studies Moscow. - 2023. - № 3. - С. 8-19. - <https://doi.org/10.17323/vo-2023-17951>
16. Абдралилов Б.С., Алтыбаева Ш.Б., Шинетова Л.Е. Результаты исследования системы ЕНТ, проведенного в рамках проекта Всемирного банка // Педагогические измерения. - 2025. - № 1(1). - С. 5-19. - <https://doi.org/10.63597/UTO3105-4161.2025.1.1.001>

Ж.Ж. Бейсенова, Л.Е. Шинетова, Г.К. Жабасева, С.Ж. Анафияева

РАШ МОДЕЛІ НЕГІЗІНДЕ ПЕДАГОГТАРДЫ ТЕСТІЛЕУДЕГІ ТЕСТ ТАПСЫРМАЛАРЫНЫҢ ДИФФЕРЕНЦИАЛДЫ ЖҰМЫС ІСТЕУІ

Мақалада педагогтерге арналған тесттің үш параллель нұсқасының материалдары негізінде жыныс белгісі бойынша тест тапсырмаларының дифференциалды жұмыс істеуін (DIF) жүргізілген психометрикалық талдау нәтижелері ұсынылған. Зерттеудің мақсаты – Раш моделі аясында DIF әсерлерінің болуын, шамасы мен бағытын бағалау, сондай-ақ көптік салыстыруларды бақылау кезінде олардың статистикалық тұрақтылығын анықтау. Параллель нұсқалардың қиындық деңгейлері

өзара салыстырмалы болған жағдайда DIF әсерлерінің үлесі мен бағыты жекелеген тапсырмалармен шектеліп, белгілі бір жыныс тобының пайдасына жүйелі ығысуды қалыптастырмайды деген болжам жасалды. Зерттеу материалы ретінде 1417 тәстіленушінің (1071 әйел және 346 ер адам) тәстілеу нәтижелері пайдаланылды. Тәстіленушілер әрқайсысы дихотомиялық тәсілмен бағаланатын 50 тапсырмадан тұратын тесттің үш нұсқасын орындады. Деректерді өңдеу Winsteps 5.8.3.0 бағдарламасында жүргізілді. Тапсырмалардың дифференциалды жұмысы DIF шамасы бойынша бағаланды және қосымша Mantel-Haenszel статистикасы арқылы тексерілді. Жалған анықтаулар үлесін (FDR) бақылау үшін әрбір тест нұсқасына жеке қолданылған Бенджамини-Хохберг процедурасы пайдаланылды. DIF-тің бастапқы критерийі бойынша «тапсырма × нұсқа» түріндегі 150 позицияның 25-інде (16,7%) DIF анықталды. FDR түзетуінен кейін статистикалық тұрғыдан тұрақты 9 позиция (6,0%) қалды. Тұрақты әсерлердің бағыты іс жүзінде теңгерімді болды: 4 тапсырма әйелдер үшін, ал 5 тапсырма ер адамдар үшін қиынырақ болды. Бұл тест тапсырмалары қиындықтарының белгілі бір жыныс тобының пайдасына жүйелі түрде ығысқанын айтуға негіз бермейді. DIF-сигналдарының ең жоғары шоғырлануы 703-нұсқада байқалды. Жалпы айтқанда, алынған нәтижелер пайдаланылған тест нұсқаларының жыныс белгісі бойынша өлшеу инварианттылығының қолайлы деңгейін көрсетеді және тесттің параллель нұсқаларын әзірлеу мен қолдану барысында DIF-ке тұрақты мониторинг жүргізу жөнінде ұсынымдар әзірлеуге мүмкіндік береді.

Түйін сөздер: тапсырмалардың дифференциалды жұмыс істеуі (DIF), Раш моделі, өлшеу инварианттылығы, педагогтарды тәстілеу, гендерлік айырмашылықтар, Mantel-Haenszel статистикасы.

Zh.Zh. Beisenova, L.E. Shinetova, G.K. Zhabayeva, S.Zh. Anafiyeva

DIFFERENTIAL ITEM FUNCTIONING IN TEACHER TESTING BASED ON THE RASCH MODEL

This article presents the results of a psychometric analysis of differential item functioning (DIF) by gender based on three parallel forms of a test for teachers. The aim of the study was to assess the presence, magnitude, and direction of DIF effects within the Rasch model and to determine their statistical robustness while controlling for multiple comparisons. It was hypothesized that, given comparable difficulty across the parallel test forms, the proportion and direction of DIF effects would be limited to individual items and would not indicate systematic bias in favor of either gender. The study was based on the test results of 1,417 examinees (1,071 women and 346 men) who completed three test forms, each consisting of 50 dichotomously scored items. Data were analyzed using Winsteps 5.8.3.0. Differential item functioning was assessed based on the magnitude of DIF, with additional testing using the Mantel–Haenszel statistic and control of the false discovery rate (FDR) using the Benjamini–Hochberg procedure, applied separately to each test form. According to the initial DIF criterion, DIF was identified in 25 of the 150 item × form positions (16.7%). After FDR correction, 9 positions (6.0%) remained statistically robust. The direction of the robust effects was nearly balanced: 4 items were more difficult for women and 5 items were more difficult for men, providing no evidence of a systematic shift in item difficulty in favor of either gender. The highest concentration of DIF signals was observed in Form 703. Overall, the results indicate an acceptable level of measurement invariance across gender for the test forms examined and provide a basis for recommendations regarding regular DIF monitoring in the development and operational use of parallel test forms.

Keywords: differential item functioning (DIF), Rasch model, measurement invariance, teacher testing, gender differences, Mantel–Haenszel statistic.

References

1. Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests. Chicago, IL: The University of Chicago Press.
2. Linacre, J. M. (2024). Winsteps® Rasch measurement computer program user's guide. Table 30.1 "Differential item functioning (DIF) pairwise." Portland, OR: Winsteps.com. Retrieved from <https://www.winsteps.com/winman/index.htm?differentialitemfunctioningdif.htm>

3. Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.
4. Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Thousand Oaks, CA: Sage Publications.
5. Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129–145). Hillsdale, NJ: Lawrence Erlbaum Associates.
6. Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719–748.
7. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
8. Mapuranga, R., Dorans, N. J., & Middleton, K. (2008). A review of recent developments in differential item functioning (ETS Research Report RR-08-43). Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/j.2333-8504.2008.tb02129.x>
9. Monahan, P. O., McHorney, C. A., Stump, T. E., & Perkins, A. J. (2007). Odds ratio, delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. Journal of Educational and Behavioral Statistics, 32(1), 92–109. <https://doi.org/10.3102/1076998606298035>
10. Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. Frontiers in Psychology, 10, Article 1137. <https://doi.org/10.3389/fpsyg.2019.01137>
11. Liu, O. L., & Wilson, M. (2009). Gender differences in large-scale math assessments: PISA trend 2000 and 2003. Applied Measurement in Education, 22(2), 164–184. <https://doi.org/10.1080/08957340902754635>
12. Susongko, P., Arfiani, Y., & Kusuma, M. (2021). Determination of gender differential item functioning in Tegal-students' scientific literacy skills with integrated science (SLiSIS) test using Rasch model. Jurnal Pendidikan IPA Indonesia, 10(2), 270–281. <https://doi.org/10.15294/jpii.v10i2.26775>
13. Ozdemir, B., & AlGhamdi, H. M. (2022). Investigating the distractors to explain DIF effects across gender in large-scale tests with non-linear logistic regression models. Frontiers in Education, 6, Article 748884. <https://doi.org/10.3389/educ.2021.748884>
14. Shear, B. R. (2020). Gender bias in test item formats: Evidence from PISA 2009, 2012, and 2015 math and reading tests. Journal of Educational Measurement, 57(3), 369–396. <https://doi.org/10.1111/jedm.12372>
15. Kardanova, E. Y., & Ivanova, A. E. (2023). Psikhometricheskie issledovaniya: sovremennyye metody i novye vozmozhnosti dlya obrazovaniya [Psychometric studies: Modern methods and new opportunities for education]. Voprosy obrazovaniya / Educational Studies Moscow, (3), 8–19. <https://doi.org/10.17323/vo-2023-17951>
16. Abdrasilov, B. S., Altybayeva, Sh. B., & Shinetova, L. E. (2025). Rezul'taty issledovaniya sistemy ENT, provedennogo v ramkakh proekta Vsemirnogo banka [Results of the study of the UNT system conducted within the World Bank project]. Pedagogikalyq ölsheuder / Pedagogicheskie izmereniya / Pedagogical Measurements, (1), 6. <https://doi.org/10.63597/UTO3105-4161.2025.1.1.001>

Сведения об авторах:

Бейсенова Жанар Жумагазиновна (автор-корреспондент) - магистр, РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, руководитель отдела науки и развития рынка, Астана, Республика Казахстан, e-mail: zhanar-beisenova@mail.ru

Шинетова Ляззат Еркековна – магистр, РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, исполняющий обязанности заместителя директора, Астана, Республика Казахстан, e-mail: shinetovalyazzat24@gmail.com

Жабаева Гульдана Куатовна – магистр, РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, ведущий научный сотрудник отдела науки и развития рынка, Астана, Республика Казахстан, e-mail: guldanzhabayeva@gmail.com

Анафияева Сулупан Жанабаевна – магистр, РГП на ПХВ «Национальный центр тестирования» МНВО РК, Астана, Казахстан, ведущий научный сотрудник отдела психометрического анализа, Астана, Республика Казахстан, e-mail: sulupaaan@gmail.com

Авторлар туралы мәлімет:

Бейсенова Жанар Жумагазиновна (автор-корреспондент) - магистр, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, ғылым және нарықты дамыту бөлімінің басшысы, Астана, Қазақстан Республикасы, e-mail: zhanar-beisenova@mail.ru

Шинетова Ляззат Ермековна – магистр, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, директор орынбасарының міндетін атқарушы, Астана, Қазақстан, e-mail: shinetovalyazzat24@gmail.com

Жабаяева Гульдана Куатовна – магистр, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, ғылым мен нарықтық дамыту бөлімінің жетекші ғылыми қызметкері, Астана, Қазақстан Республикасы, e-mail: guldanzhabayeva@gmail.com

Анафияева Сулупан Жанабаевна – магистр, ҚР ҒЖБМ «Ұлттық тестілеу орталығы» ШЖҚ РМК, психометрикалық талдау бөлімінің жетекші ғылыми қызметкері, Астана, Қазақстан Республикасы, e-mail: sulupaaan@gmail.com

Information about authors:

Beisenova Zhanar Zhumagazinovna (corresponding author) – Master’s degree, Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Head of the Department of Science and Market Development, Astana, Republic of Kazakhstan, e-mail: zhanar-beisenova@mail.ru

Shinetova Lyazzat Ermekovna – Master’s degree, Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Acting Deputy Director, Astana, Republic of Kazakhstan, e-mail: shinetovalyazzat24@gmail.com

Zhabayeva Guldana Kuvatovna – Master’s degree, Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Leading Researcher, Department of Science and Market Development, Astana, Republic of Kazakhstan, e-mail: guldanzhabayeva@gmail.com

Anafiyeva Sulupan Zhanabayevna – Master’s degree, Republican State Enterprise on the Right of Economic Management «National Testing Center» of the Ministry of Science and Higher Education of the Republic of Kazakhstan, Leading Researcher, Department of Psychometric Analysis, Astana, Republic of Kazakhstan, e-mail: sulupaaan@gmail.com