УДК 316.62

АНАЛИЗ DIF В ОЦЕНКЕ ОБЩЕГО ИНТЕЛЛЕКТА ДЛЯ ГЕНЕРИРУЕМЫХ КОМПЬЮТЕРОМ ГРАФИЧЕСКИХ ТЕСТОВЫХ ЗАДАЧ В ДВУХ ЭТНИЧЕСКИ РАЗЛИЧНЫХ ВЫБОРКАХ

Ф.А. Фройнд [1], ***С.В. Давыдов*** [2], ***Й.П. Бертлинг*** [3], ***Х. Холлинг*** [3], ***Г.С. Шляхтин*** [2]

[1] Университет Оснабрюка, Германия
[2] Нижегородский госуниверситет им. Н.И. Лобачевского
[3] Вестфэлише Вильгельмс Университет Мюнстера, Германия

sgs-nn@mail.ru

Поступила в редакцию 22.08.2012

Графические матричные задания нашли широкое применение для измерения общего интеллекта. Данный тип задач в силу их невербального характера считается относительно свободным от культуры или, по меньшей мере, не столь культурно нагруженным, как, например, при использовании вербальных заданий. В то же время последние исследования показывают, что большинство невербальных тестов в той же степени культурно обусловлены, как и вербальные тесты. В данном исследовании для конструирования матричных заданий использовались алгоритмы автоматического генерирования задач. Перед выполнением тестовых заданий всем испытуемым была дана подробная инструкция с целью создания равных экспериментальных условий. Данные, полученные на немецкой и русской выборках, были проанализированы на наличие дифференцированного функционирования заданий (DIF). Было показано отсутствие DIF в каждом из заданий, что позволяет сделать вывод о том, что свободные от культуры матричные задания могут быть сконструированы при использовании принципов автоматического генерирования задач и после создания равных экспериментальных условий.

Ключевые слова: графические матричные задания, свободные от культуры тесты интеллекта, дифференцированное функционирование заданий.

Введение

Общий интеллект, часто рассматриваемый как синоним фактора «g» Спирмена – подвижного интеллекта, абстрактного или индуктивного мышления [1, 2, 3], является одним из наиболее значимых психологических параметров в плане прогнозирования эффективности профессиональной деятельности [4, 5]. Для измерения уровня развития общего интеллекта существует огромное количество различных тестов, в то же время не все эти инструменты оценивают его в равной степени успешно. Во многом это связано с тем, что итоговые результаты тестирования оказываются обусловленными не только собственно интеллектом, но и «сторонними» по отношению к нему факторами, в том числе такими, как принадлежность человека к определенной группе – культурной, этнической, гендерной и др., которая определяет сложившиеся у него алгоритмы, стереотипы, стиль, опыт и навыки решения интеллектуальных задач. Это делает прямое сравнение результатов различных тестов у одних и тех же испытуемых или одного и того же теста у разных испытуемых весьма проблематичным. Решение этой проблемы заключается в определении доли влияния на полученный результат тестирования собственно интеллекта – т.е. насыщения g-фактором (фактором общего интеллекта) конкретного теста или задания [6].

Свободные от культуры тесты интеллекта

Концепция так называемого «свободного от культуры» интеллектуального тестирования берет начало с работ Кеттелла [7, 8], разработавшего графические задания для измерения интеллекта. При помощи таких заданий Кеттелл пытался измерить независимый от культуры интеллект, опираясь на абстрактные стимулы, не связанные с каким-либо специфическим культуральным контекстом и вследствие этого не дающие преимущества одной из групп перед другой [8]. Такое преимущество может являться результатом прежнего опыта, имеющегося у носителей данной культуры, но не членами других культурных групп. Свободные от культуры тесты интеллекта были встречены весьма положительно, но, к сожалению, тесты Кеттелла вызвали те же проблемы кросскультурного сравнения, что и прочие тесты, специально не разрабатываемые для оценки свободного от куль-

туры интеллекта [9, 10, 11, 12]. Таким образом, хотя сама идея разработки свободных от культуры тестов, не включающих вербальное содержание и рассчитывающих исключительно на графический материал, правильна, она не обязательно приводит к устранению культурных влияний. Можно предположить, что члены разных этнических групп не знакомы с такими заданиями в равной степени. Когда конструируются такого рода задания, то представления об определенных (графических) стимулах и когнитивных схемах, в которые они включены, играют важную роль. Весьма явно роль культурных различий проявляется в случае интеллектуального тестирования. В том случае если полученные результаты используются для определения способностей соискателя при приеме на работу, поступлении в университет и т.д., подобные различия могут приводить к проблемам в многонациональных обществах. П. Гринфилд эксплицитно связывала интерпретацию результатов тестов на интеллект со специфическим культуральным контекстом [13]. По ее мнению, интеллектуальные тесты не могут быть использованы в качестве универсальных прикладных измерительных инструментов, поскольку они разрабатывались в рамках отдельной культурной среды и, тем самым, отражают принятые и разделяемые ценности, стратегии коммуникации и знания, превалирующие в данном контексте. Кроме того, П. Гринфилд придавала особое значение важности деликатной интерпретации результатов теста и указывала на то, что сравнение между группами имеет смысл только в том случае, если в отношении обеих групп будет показано, что измеряется один и тот же конструкт, что и при выполнении рукописного варианта теста. Проблемами интерпретации результатов теста в различных культурах также занимались Р. Серпелл и Б. Хейнес, рекомендовавшие полностью отказаться от применения тестов интеллекта в том случае, если их результаты используются для профотбора, поскольку имеющиеся в настоящее время в распоряжении психологов измерительные инструменты в целом не доказали своей независимости от культуры [14]. Важно указать, что эти выводы применимы не только в отношении исключительно культурно противоположных контекстов, но и в отношении культур с намного более тесными связями. Например, нельзя гарантировать, что тест, разработанный в Германии и перенесенный в Россию, является свободным от культуры, поскольку он основан исключительно на графическом материале, предполагающем его универсальность. Предположения, подобные этому, следует всегда тщательно проверять, поскольку возможно, что определенные задания дают преимущества какой-то одной культурной группе и при этом в явном виде не обнаруживаются.

Дифференцированное функционирование заданий

Искажения в тестовых заданиях могут быть выявлены при помощи процедуры дифференцированного функционирования заданий (DIF). DIF определяется как поддающееся количественному определению различие в измеряемых характеристиках тестового задания для двух или более групп [15, 16]. В DIF-анализе для двух групп сравниваются референтная и фокусная группы. Для анализа DIF существует большое разнообразие статистических процедур, среди которых можно выделить методы теории тестовых заданий – Item Response Theory (сокращенно – IRT [см., напр., 17]), регрессионного анализа (logistic regression (напр**.,** Swaminathan & Rogers, 1990), и методы таблиц сопряженности (contingency-table methods; [18]). Важное различие может быть проведено в отношении стандартных и нестандартных DIF. Стандартные DIF имеют место, когда уровень DIF не зависит от уровня способностей референтной и фокусной групп, в то время как нестандартные DIF существуют тогда, когда коэффициент вероятности правильно решенного задания для референтной группы отличается от коэффициента вероятности для фокусной группы в различных точках на спектре способностей [19]. В контексте IRT, стандартные DIF указывают на межгрупповые различия, обусловленные сложностью тестового задания, а нестандартные DIF указывают на межгрупповые различия в разрешающей способности тестового задания. Если модель Раша соответствует данным**,** тогда оценка и стандартных и нестандартных DIF может быть рассмотрена как дополнительный критерий адекватности теста.

Графические матричные задания

Графические матричные задания уже давно используются в интеллектуальном тестировании. Данная категория заданий считалась одной из лучших процедур для измерения фактора *g* [6, 20]. Матричные задания являются полностью невербальными. На рис. 1 показано, что они составлены из графических элементов, подчиняющихся определенным логическим правилам, (напр., вращения объекта – см. большой черный пятиугольник, который вращается против часовой стрелки по рядам матрицы), которые должны быть правильно идентифицированы и использованы для решения задачи. Как правило, правый нижний элемент матрицы оставлен пустым для записи верного решения, которое должно быть выбрано из набора дистракторов.

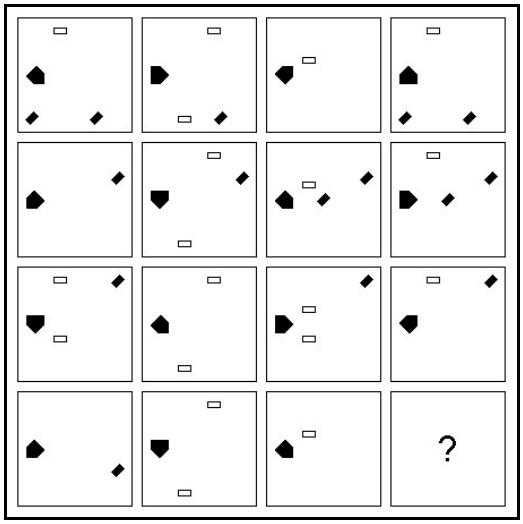

Рис. 1. Пример матричного тестового задания

Наиболее известными матричными тестами являются так называемые «прогрессивные матрицы» (PM) Дж. Равена, насчитывающие 3 формы: стандартный вариант (SPM), продвинутый вариант (APM), разработанный для лиц с повышенными интеллектуальными способностями, и вариант для детей (CPM). Вследствие своего невербального характера матричные тесты нашли широкое применение по всему миру и считались благодаря работам Кеттелла культурно-независимыми. Большинство исследований, изучавших данный вопрос, были проведены еще в прошлом веке [21, 22, 23] . В общем и целом, было установлено, что матричные тесты являются относительно объективными в отношении различных культур. Несмотря на это приведенные доказательства далеко не очевидны. Убедительным было бы только полное отсутствие отклонений во всех тестовых заданиях, что экспериментально доказано не было.

DIF-методы уже использовались в тестировании для выявления различий в психометрических характеристиках матричных заданий между мужчинами и женщинами [24]. Эти исследования вновь подтвердили наличие дискриминирующих отклонений в некоторых из тестовых заданий Равена, как правило в пользу мужчин. Исходя из этого, представляется важным показать, что матричные задания могут найти применение для измерения общего интеллекта в различных культурах при использовании соответствующих методов DIF-анализа.

Автоматическое генерирование задач и равные экспериментальные условия

Зачастую разработчикам тестовых заданий для их создания приходится работать, по сути, вручную, применяя свой уникальный авторский стиль, что превращает их работу в настоящее «искусство». Им приходится принимать множество решений, и каждое задание подвергается процедуре серьезного контроля качества. В противоположность этому, принципы автоматического генерирования задач дают конструкторам тестов возможность программного создания высококачественных тестовых заданий на основе конструктивных логических обоснований, связанных с когнитивными теориями, которые могут рассматриваться как универсальные [25, 26]. Эти конструктивные логические обоснования могут быть заложены в компьютеризированные алгоритмы, и если предполагается, что они универсальны, то это то предположение, которое может быть проверено при проведении DIF-анализа.

Другое преимущество автоматического генерирования задач состоит в адаптивном тестировании, поскольку для реализации наиболее точного измерения способности необходимо наличие обширных наборов тестовых заданий [27]. В случае использования методов автоматического генерирования задач могут генерироваться параллельные тестовые задания с ожидаемыми психометрическими характеристиками [28]. Это позволяет обеспечить надежность теста, которая является важным компонентом любых видов тестов.

Наряду с надежностью теста, может быть также повышена и test fairness[1] путем предоставления достаточной информации непосредственно перед прохождением теста. А. Анастази отсылает к этому способу инструктирования как к краткому тестоориентировочному заданию [29]. Большинство тестов на интеллект, среди них и прогрессивные матрицы, предлагают несколько тренировочных заданий, но они не дают информации о принципах их конструирования. Таким образом, хотя на первый взгляд предоставление тренировочных заданий может показаться достаточным для создания равных экспериментальных условий, однако это не гарантирует того, что члены различных групп (этнических, гендерных, социоэкономических и др.) обрабатывают содержание задания одним и тем же способом. Вполне возможным результатом этого оказывается систематическая ошибка задания, отображаемая при помощи DIF-анализа.

Настоящее исследование является, насколько мы знаем, первым изучающим культуральные DIF в автоматически генерируемых графических матричных заданиях. Для анализа были взяты данные, полученные на двух этнических группах – немецких и русских студентов. Несмотря на то что фактически ни одно из проведенных до настоящего времени исследований

не показало, что матричные задания являются полностью свободными от влияния культуры, мы ожидали, что в данном исследовании в автоматически генерируемых заданиях будет выявлено отсутствие DIF. Наше предположение было обусловлено следующим: а) конструктивное логическое обоснование базировалось на претендующей на универсальность когнитивной теории [20, 23, 30], б) непосредственно перед прохождением теста его участникам была дана детальная инструкция с целью обеспечения равных стартовых условий и в) на содержание заданий не оказывал влияния авторский стиль создателей теста.

Метод

Характеристика выборки

Суммарная выборка состояла из 152 студентов. Референтная группа состояла из 94 немецких студентов, а фокусная группа из 58 русских студентов. 91 испытуемый ранее проходил тестирование на интеллект, 54 – нет и 7 – не дали явного ответа. Средний возраст выборки составил около 21 года, однако средний возраст в двух подвыборках значительно различался: студенты русской подвыборки были в среднем на 5 лет моложе (t [103.036] = 10.929, p = .000). Причиной этого являются различия в системах университетского образования двух стран. В России студенты оканчивают школу в гораздо более юном возрасте, чем в Германии и, следовательно, раньше поступают в университет. Кроме того, дополнительный возрастной сдвиг в немецкой подвыборке обусловлен тем фактом, что некоторые студенты поступили в университет после нескольких лет ожидания, поскольку не имели права обучаться по выбранной специальности сразу же после окончания школы вследствие низких средних баллов в аттестате. Что касается гендерного аспекта, то в совокупной выборке преобладали женщины в отношении 3.6 : 1. В таблице 1 приведена более детальная информация по выборке.

Материал

Для генерирования набора матричных заданий (k = 15) использовалась программа *MatrixDeveloper* [31]. В каждом задании присутствовало до 5 различных правил. Задания были представлены в формате multiple choice. Каждое из заданий имело 17 вариантов ответа, где 17-м вариантом было «верное решение отсутствует» с целью уменьшения попыток угадывающего поведения и устранения стратегии фальсификации. Перед началом теста каждому из участников теста был роздан отдельный инструкционный буклет, в котором были объяснены все конструкционные параметры с целью снижения потенциального преимущества от ранее полученного опыта выполнения тестов и для обеспечения максимальной test fairness [28]. Все материалы были представлены в бумажном виде. Тестирование проводилось в группах с соблюдением строгих условий без лимита времени. Мотивация участников к тестированию оценивалась при помощи краткого опросника QCM [32]. Интерес, вызов, ожидаемая вероятность успеха и опасение неудачи оценивались по 7-балльной шкале.

Статистические процедуры

Для совокупной выборки была использована модель Раша [33]. Из большого числа различных имеющихся показателей [34] в настоящем исследовании был выбран Q-индекс [35]. Для проведения DIF-анализа по причине достаточно небольшого размера выборок были использованы непараметрические методы. Анализ стандартного DIF проводился с использованием хорошо известного критерия Мантеля – Гензеля χ^2-test [18, 36]. Для оценки нестандартного DIF использовался критерий Бреслоу – Дэя [37, 19], обозначаемый как B-D χ^2. Его мощность относительно высока при средней сложности исследуемого задания, но падает при явном повышении ее уровня [19]. Также нами было применено «Комбинированное правило принятия решения», обозначенное как CDR. Оно проверяет нулевую гипотезу на отсутствие DIF в задании, если оба критерия – M-H χ^2 и B-D χ^2 – приводят к решению принятия нулевой гипотезы, и отбрасывает предположение об отсутствии DIF, если какой-то из двух критериев – M-H χ^2 или B-D χ^2 – указывает на его наличие. Поэтому каждое задание анализировалось на наличие стандартных и нестандартных DIF. Наконец, классификационная схема ETS (службы образовательного тестирования) выявляет уровень DIF в задании на основании трех категорий (A, B, C), где категория A обозначается как не имеющая или имеющая незначительный DIF, B – как показывающая средний уровень DIF и C – как показывающая высокий DIF [38]. Схема классификации ETS также использует критерий M-H χ^2.

Результаты

Среднее время обработки 15 тестовых заданий составило 53 минуты (SD = 13мин). Между

немецкими и русскими испытуемыми не было выявлено значимых различий в среднем времени выполнения заданий (t [78.616] = – 1.128, p = .263). Средний сырой балл для совокупной выборки составил 7.78, но он был выше в немецкой выборке, чем в русской (t [150] = 2.532, p = .012). Внутренняя согласованность тестовых баллов была очень хорошей (α = .89).

В целом, все участники рассматривали тест как вызов – проявляли интерес и оценивали ожидаемую вероятность успеха несколько выше среднего, а вероятность неудачи – ниже среднего. Единственное значимое различие между группами было выявлено по показателю ожидаемой вероятности успеха (F [1,151] = 11.501, p = .001), которая у русских испытуемых была намного выше. В таблице 1 приведены некоторые данные описательной статистики.

Согласно данным совокупной выборки, Q-индекс показал, что модель Раша хорошо совместима со всеми 15 заданиями. Параметры сложности заданий были стандартизированы по сумме и находились в диапазоне между –0.611 и 0.566, указывая скорее на среднюю сложность для всех заданий в наборе. В таблице 2 дается детальная информация о параметрах сложности задания и его пригодности.

Анализ на наличие DIF проводился с использованием вышеописанных процедур. Таблица 3 показывает, что из 15 заданий в одном из них под № 4 было выявлено нестандартное DIF. Поскольку ни одно из заданий не оказалось отбракованным в результате проведения М-Н χ^2 и B-D χ^2 анализа, то на основании комбинированного правила принятия решения делается вывод об отсутствии значимых DIF в каждом из 15 заданий. Этот вывод был подкреплен схемой классификации DIF службы образовательного тестирования ETS.

Обсуждение

Результаты DIF-анализа подтвердили наше ожидание, что рациональный подход к конструированию заданий, совмещенный с детальным предварительным пояснением, способен

Таблица 1

Характеристики выборки

	M возраст	Станд.откл	Пол		M время	Станд.откл	M сырой балл	Станд.откл	M интерес	Станд.откл	M вызов	Станд.откл	M вероятность успеха	Станд.откл	M опасение неудачи	Станд.откл
			ж	м												
Совокупная выборка	21	4.44	119	33	53.6	12.9	7.78	4.63	4.14	1.35	5.00	1.06	4.41	1.10	3.32	1.37
Немецкая подвыборка	23	4.55	72	22	52.5	9.4	8.51	4.62	3.99	1.34	4.99	1.11	4.18	1.06	3.26	1.37
Русская подвыборка	18	0.83	47	11	55.3	17.1	6.59	4.44	4.37	1.33	5.00	0.98	4.79	1.07	3.43	1.37

Таблица 2

Данные IRT параметров

Задание	β_i^*	*S.E.*	Z_Q
01	-0.451	.207	0.069
02	-0.172	.207	-0.324
03	-0.092	.207	0.561
04	0.273	.210	-0.515
05	0.314	.211	-0.025
06	-0.252	.207	0.651
07	-0.451	.207	0.478
08	0.232	.210	-0.662
09	0.566	.214	-0.522
10	-0.611	.208	-0.412
11	-0.252	.207	0.024
12	-0.052	.208	0.328
13	0.356	.211	-0.442
14	0.566	.214	0.144
15	0.028	.208	0.609

* - β_i = параметр сложности задания; *S.E.* = среднеквадратическая погрешность параметра сложности задания; Z_Q = Q-индекс пригодности задания.

Таблица 3

DIF статистика

Задание	M-H χ^2	B-D χ^2	CDR	ETS
01	1.617	0.995	OK	A
02	0.393	0.162	OK	A
03	1.139	1.444	OK	A
04	0.155	4.917*	OK	A
05	0.661	0.277	OK	A
06	0.029	0.729	OK	A
07	2.245	0.057	OK	A
08	0.558	2.358	OK	A
09	0.052	0.117	OK	A
10	0.018	3.553	OK	A
11	0.002	1.071	OK	A
12	0.023	0.055	OK	A
13	0.668	2.312	OK	A
14	0.009	0.389	OK	A
15	2.127	0.634	OK	A

* - $p < .05$; M-H χ^2 = Mantel-Haenszel χ^2; B-D χ^2 = Breslow-Day χ^2, CDR = комбинированное правило принятия решения; ETS = служба образовательного тестирования (схема классификации DIF).

полностью нивелировать влияние культуры в графических матричных заданиях. Этот обнадеживающий результат показывает, что, опираясь на универсальные конструкционные принципы и на тщательную инструкцию, вполне возможно создавать равные экспериментальные условия. Некоторые детали настоящего исследования заслуживают более пристального внимания.

Во-первых, важнейшим отличием от других исследований, изучавших культурные (или какие-либо другие) влияния в матричных заданиях, является предоставление поясняющих раздаточных материалов. Эти материалы дают значительно больше, чем просто предоставление нескольких образцов тренировочных заданий. В действительности это удовлетворяет всем требованиям, предъявляемым А. Анастази, предлагавшей обеспечивать всем тестируемым равные условия прохождения теста, независимо от того, имелся ли у них прежний опыт такого тестирования, дающий определенное преимущество над теми, у кого такого опыта нет [29]. Предшествующие исследования показали, что наличие опыта тестирования помогает даже в том случае, когда используются различные типы заданий, поскольку имеет место перенос навыков тестирования [39]. Значительная часть совокупной выборки (60%) уже имела опыт прохождения интеллектуальных тестов, однако благодаря наличию раздаточных материалов данный факт не дал им статистически значимых преимуществ перед другими. Необходимость предоставления инструктивного раздаточного материала также была продиктована результатами более раннего исследования [40], показавшего наличие «эффекта пола»[2] (floor effect) в сырых баллах заданий при отсутствии инструкций. К тому же, как показал анализ Раша на совокупной выборке и на обеих подвыборках, в тесте отсутствовали очень легкие задания. В совокупной выборке ни одно из заданий не было решено более чем 62% участников. По этой причине мы не стали экспериментально манипулировать фактором «инструкция», добавив контрольную группу, не получившую раздаточных инструктивных материалов. Можно сделать вывод, что инструкция имела смысл, поскольку средний уровень сложности задания был достигнут.

Во-вторых, необходима весьма осторожная оценка различий, полученных в средних сырых баллах в немецкой и русской подвыборках. Немецкая подвыборка состояла преимущественно из студентов-психологов в возрасте 23 лет. В Германии студент, желающий изучать психологию в университете, должен иметь очень высокий проходной аттестационный средний балл, т.е. быть в первых рядах выпускников гимназии как высшего уровня школьного образования. По сравнению с этим, русские студенты были в среднем на 5 лет моложе своих немецких визави в референтной группе, да и сам состав испытуемых русской выборки был не столь академически успешен.

Поскольку гендерные влияния в матричных заданиях изучались ранее и имели неоднозначные результаты [24], было весьма интересно проверить, способен ли подход к тестированию, который был применен в настоящем исследовании, также устранить DIF между испытуемыми мужского и женского пола. К сожалению, для проведения DIF с использованием пола в качестве критериальной переменной в выборке явно недоставало представителей мужского пола. По мнению Ф. Абад и др. (Abad et al., 2004) мат-

ричные тесты дают преимущество мужчинам, когда некоторые задания имеют визуально-пространственную природу. Это может быть справедливым в отношении матриц Равена, однако автоматически генерируемые задания, использованные в настоящем исследовании, не используют каких-либо 3D-эффектов, например таких, как соединение частей фигур, следовательно, отсутствует и визуально-пространственное содержание. Следует констатировать, что организация настоящего исследования не позволила нам выявить гендерные DIF, однако это та проблематика, которая обязательно должна быть адресована будущим исследованиям.

Примечания

1. При тестировании всем испытуемым должны быть предоставлены одинаковые инструкции, задания, условия, правила интерпретации и оценивания результатов, одинаковое время и т.д. Все эти и другие требования этического характера относятся к словосочетанию test fairness.

2. Эффект пола (floor effect) отражает чрезмерную трудность задания, так что многие испытуемые вообще не могут с ним справиться.

Список литературы

1. Cattell R.B., Horn J.L. A check on the theory of fluid and crystallized intelligence with description of new subtest designs // Journal of Educational Measurement. 1978. 15. 139–164.

2. Court J.H. Raven Progressive Matrices // In R. J. Sternberg (Ed.), Encyclopedia of human intelligence. New York: Macmillan, 1994. P. 916–917.

3. Spearman C. The abilities of man. New York: Macmillan, 1927.

4. Ones D.S., Viswesvaran C., Dilchert, S. Cognitive abilities in selection decisions // In O. Wilhelm & R. W. Engle (Eds.), Handbook of understanding and measuring intelligence. Thousand Oaks, CA: Sage Publications, 2005. P. 431–468.

5. Schmidt F.L., Hunter J.E. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings // Psychological Bulletin. 1998. 124. P. 262–274.

6. Bartholomew D.J. Measuring intelligence. Facts and fallacies. Cambridge, UK: Cambridge University Press, 2004.

7. Cattell R.B. A culture free intelligence test I // Journal of Educational Psychology. 1940. 31. P. 161–180.

8. Cattell R.B. Are culture fair intelligence tests possible and necessary? // Journal of Research and Development in Education. 1979. 12. P. 2–13.

9. Cole M. Cultural Psychology. A once and future discipline. Cambridge, MA: Belknap Press, 1996.

10. Nenty H. Cross-culture bias analysis of Cattell Culture-Fair Intelligence Test // Perspectives in Psychological Researches. 1986. 9. P. 1–16.

11. Nenty H.J., Dinero T.E. A cross-cultural analysis of the fairness of the Cattell Culture Fair Intelligence Test using the Rasch model // Applied Psychological Measurement. 1981. 5. P. 355–368.

12. Ross N. Culture & Cognition. Implications for theory and method. Thousand Oaks: Sage, 2004.

13. Greenfield P.M. You can't take it with you – why ability assessments don't cross cultures // American Psychologist. 1997. 52. P. 1115–11224.

14. Serpell R., Haynes B.P. The cultural practice of intelligence testing: Problems of international export // In R. J. Sternberg & E. L. Grigorenko (Eds.), Culture and competence. Washington, DC: American Psychological Association, 2004. P. 163–185.

15. Camilli G., Shepard L.A. Methods for identifying biased test items. Newbury Park, CA: Sage, 1994.

16. Dorans N.J., Holland P.W. DIF detection and description: Mantel-Haenszel and standardization // In P.W. Holland, & H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Earlbaum, 1993. P. 35–66

17. Lord F.M. Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum, 1980.

18. Holland P.W., Thayer D.T. Differential item performance and the Mantel - Haenszel procedure // In H. Wainer & H. I. Braun (Eds.), Test validity. Hillsdale, NJ: Lawrence Erlbaum, 1988. P. 129–145

19. Penfield R.D. Application of the Breslow-Day test of trend in odds ratio heterogeneity to the detection of nonuniform DIF // Alberta Journal of Educational Research. 2003. 49. P. 231–243.

20. Carpenter P.A., Just M.A., Shell, P. What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices test // Psychological Review. 1990. 97. P. 404–431.

21. Bart W.M., Rothen W., Read S. An ordering-analytic approach to the study of group differences in intelligence // Educational and Psychological Measurement. 1986. 46. P. 799–810.

22. Corman L., Budoff M. Factor structures of Spanish-speaking and non-Spanish-speaking children on Raven's Progressive Matrices // Educational and Psychological Measurement. 1974. 34. P. 977–981.

23. Jensen A.R. How biased are culture-loaded tests? // Genetic Psychology Monographs. 1974. 90. P. 185–244.

24. Abad F.J., Colom R., Rebollo I., Escorial S. Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias // Personality and Individual Differences. 2004. 36. P. 1459–1470.

25. Arendasy M., Sommer M., Gittler G., Hergovich A. Automatic Generation of Quantitative Reasoning Items: A Pilot Study // Journal of Individual Differences. 2006. 27. P. 2–14.

26. Embretson S.E. A cognitive design system approach to generating valid tests: Application to abstract reasoning // Psychological Methods. 1998. 3. P. 380–396.

27. Van der Linden W.J. Linear models for optimal test design. New York: Springer, 2005.

28. Freund Ph.A., Hofer S., Holling H. Explaining and controlling for the psychometric properties of computer-generated figural matrices items // Applied Psychological Measurement. 2008. 32. P. 195–210.

29. Anastasi A. Coaching, test sophistication, and developed abilities // American Psychologist. 1981. 36. P. 1086–1093.

30. Preckel F. Diagnostik intellektueller Hochbegabung. Testentwicklung zur Erfassung der fluiden Intelligenz [Assessment of intellectual giftedness: Test development for the assessment of fluid intelligence]. Göttingen: Hogrefe, 2003.

31. Hofer S. Matrix Developer. Unpublished computer software, University of Münster, 2006.

32. Rheinberg F., Vollmeyer R., Burns B.D. FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern- und Leistungssituationen [QCM: A questionnaire for the assessment of current motivation in learning situations] // Diagnostica. 2001. 47. P. 57–66.

33. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Educational Research Institute, 1960.

34. Glas C.A.W., Verhelst N.D. Testing the Rasch model // In G. H. Fischer & I. W. Molenaar (Eds.), Rasch models. Foundations, recent developments, and applications. New York: Springer, 1995. P. 69–95.

35. Rost J., von Davier M. A conditional item fit index for Rasch models // Applied Psychological Measurement. 1994. 18. P. 171–182.

36. Mantel N., & Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease // Journal of the National Cancer Institute. 1959. 22. P. 719–748.

37. Breslow N.E., Day N.E. Statistical methods in cancer research: Volume 1 – The analysis of case-control studies. Lyon: International Agency for Research on Cancer, 1980.

38. Zieky M. Practical questions in the use of DIF statistics in item development // In P. W. Holland & H. Wainer (Eds.), Differential item functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993. P. 337–364.

39. Nevo B. The effects of general practice, specific practice, and item familiarization on change in aptitude test scores // Measurement and Evaluation in Guidance. 1976. 9. P. 16–20.

40. Müller A. Variation Einfachbezug versus Mehrfachbezug und Training versus kein Training bei regelgeleitet konstruierten Matrizenaufgaben auf die Ergebnisleistung [Variation of single vs. multiple rules per element and training vs. no training in rule-based generated matrix items and performance]. Unpublished master's thesis, University of Muenster, 2001.

41. Holland P. W., & Wainer H. Differential Item Functioning. Hillsdale, NJ: Lawrence Erlbaum, 1993.

42. Raven J. C. Advanced Progressive Matrices, set II. London: H. K. Lewis, 1962.

43. Swaminathan H., & Rogers H. J. Detecting differential item functioning using logistic regression procedures // Journal of Educational Measurement. 1990. 27. P. 361–370.

DIF ANALYSIS IN THE ASSESSMENT OF GENERAL INTELLIGENCE FOR COMPUTER-GENERATED GRAPHICS TEST PROBLEMS IN TWO ETHNICALLY DISTINCT SAMPLES

P F.A. Freund, S.V. Davydov, J.P. Bertling, H. Holling, G.S. Shlyakhtin

Figural matrix items are widely used for measuring general intelligence. Because of its non-verbal character, this task type is considered to be relatively culture-fair, or at least not as culture-loaded as, for instance, test items with verbal content. However, recent investigations have shown that most non-verbal tests are as much culturally biased as verbal tests. In this study, automatic item generation algorithms were employed to construct matrix items. Test takers received a thorough instruction before the actual test in order to facilitate equal testing conditions. Data from a German and a Russian sample were investigated for the presence of Differential Item Functioning (DIF). It could be shown that there was no DIF in any of the items, leading to the conclusion that culture-fair matrix items can be generated using principles of automatic item generation and after creation of equal testing conditions.

Keywords: figural matrix items, culture-fair intelligence tests, differential item functioning.