Приложимост на Класическата тестова теория и Теорията за отговор на тестов въпрос. Преглед на литературата по въпроса

Л. Джалев

Нов български университет, София, България

Резюме

Целта на този литературен обзор е да се представи състоянието на изследванията върху приложимостта на Класическата тестова теория и Теорията за отговор на тестов въпрос, главно върху съответствието между основните допускания за нормалност на разпределенията на латентните променливи, едномерност на латентното пространство и локална независимост на отговорите, и профила на данните в областта на скалирането на когнитивни способности (постижения), както и някои други характеристики на двете теории.

Обзорът е част от едно по-широко изследване, проведено в контекста на засиления интерес към теоретичните и приложните аспекти на психологическите измервания в поведенческите, социалните и хуманитарните науки. То е фокусирано върху две психометрични теории, които, без съмнение, са доминиращи в тази научна област. Теориите притежават две важни характеристики, от които произтичат основните направления в настоящото изследване: (а) моделите, разработени в техните рамки, са модели на данни и (б) те са функционално еквивалентни, но независими, алтернативни една на друга. Тези особености предполагат необходимостта от проучване на приложимостта на теоретичните

тестови модели не само от общометодологически интерес, но и при всяко конкретно изследване.

За съжаление, проучванията по този въпрос са твърде малко и, доколкото ни е известно, няма нито едно, в което тази проблематика да е обхвана в нейната пълнота и дълбочина. По-голяма част от резултатите, които се отнасят предимно до IRT, са страничен продукт в значително по-обширната специализирана литература, посветена на параметричните статистики и на тяхната устойчивост. В допълнение, приложението на тестовите теории в много области извън психологическите измервания на постижения поставя различни от поставените, но не по-малко интересни въпроси. Поради това в обзора са представени и някои изследвания на приложимостта на тези теории извън заявената предметна област.

Резултатите от изследванията, фокусирани конкретно върху приложимостта на тестовите модели, както и онези с други изследователски задачи, но с използване на един или друг тестов модел, показват по-скоро отсъствие на условията за тяхната приложимост.

Ключови думи: *Класическа тестова теория*, *Теория за отговор на тестов въпрос*, *теория*, *модел*, *приложимост*, *латентна променлива*, *нормално разпределение*, *едномерност*, *локална независимост*

1. Въведение

Представеният литературен обзор е част от едно по-широко изследване, проведено в контекста на засиления интерес към теоретичните и приложните аспекти на психологическите измервания в поведенческите, социалните и хуманитарните науки и, през последните 15 години - в медицинските науки. То е фокусирано върху две психометрични теории, които, без съмнение, са доминиращи в тази научна област: Класическата тестова теория (*Classical test theory,* CTT) и Теорията за отговор на тестов въпрос (*Item response theory,* IRT). Теориите притежават две важни характеристики, от които произтичат основните направления в това изследване: (а) моделите, разработени в техните рамки, са модели на данни и (б) тези теории са функционално еквивалентни, но независими, алтернативни една на друга и тяхното приложение за решаване на една или друга научна или приложна задача е въпрос на обоснован избор от страна на изследователя.

Основно поле на приложение на двете психометрични теории са образователните тестове, какъвто е Тестът по общообразователна подготовка (ТОП). Това е първата мащабна, устойчива и практически непроменена тестова програма за конструиране и администриране на инструменти за психологически измервания в нашата образователна система, която стои в основата на приемните процедури в Нов български университет от 1996 година насам. Поради особеностите на кандидатстудентската кампания в НБУ, чиято продължителност е 8 месеца и през която се провеждат 5 изпитни сесии с 20 – 22 изпитни варианта на теста, в продължение на повече от 15 години е натрупана огромна база от емпирични данни. Методологията за конструиране и прилагане на теста, както и за обработване и интерпретация на получените резултати, към настоящия момент е изцяло в рамките на Класическата тестова теория. Това обстоятелство повдига два важни въпроса: (а) оправдано ли е прилагането на Класическата тестова теория и (б) допустимо ли е прилагането на Теорията за отговор на тестов въпрос.

2. Необходимост от изследване на приложимостта

Преди да представим аргументите в полза на необходимостта от изследване на приложимостта, е необходимо да разгледаме съдържанието на понятията „(тестова) теория" и „(тестов) модел" и свързаното с тях понятие „приложимост". Теорията представлява обща, абстрактна концептуална рамка на определен сегмент от наблюдаваната реалност. Тестовите теории включват понятия от по-високо равнище като наблюдавани и латентни променливи, съответно наблюдаван и действителен бал, трудност на въпроса или неговата характеристична крива, както и отношенията (връзките) между тях. Поради обобщения характер на понятията и дефинираните връзки между тях, една тестова теория не следва да се оценява от гледна точка на нейната полезност (Hambleton & Jones, 1993).

Моделът представлява конкретизация на една или друга теория, със специфични определения на теоретичните понятия и техните връзки, предназначен за моделиране на отделни феномени или група от сродни феномени, които са част от полето на съответната теория. Тестовите модели се изграждат в рамките на определена тестова теория и описват по-конкретно, точно и подробно както понятията, така и връзките между тях, обикновени под формата на математически изрази. Обикновено моделите включват и определен набор от допускания относно съдържанието и/ или характера на понятията и връзките между тях, които могат да се разглеждат като *условия* за приложението на съответния модел. Поради това един теоретичен модел следва да се проверява от гледна точка на неговото съответствие с реалните феномени, към които ще бъде приложен. При тестирането оценката на годността на съответния модел следва да се прави върху конкретна съвкупност от тестови данни, с добре обмислена система от емпирични методи.

Тестовите теории и съпътстващите ги модели са от изключителна важност за практиката на образователните и психологическите измервания, защото предлагат рамка за третирането на редица важни проблеми като грешката на измерване, типа на връзката между способностите и тестовите въпроси и редица други проблеми, позволяващи конструирането на тестове с предварително зададени и желани характеристики.

Психометричните тестови модели могат да бъдат класифицирани като „модели на данни” (Suppes, 1962; Frigg & Hartmann, 2006). Описвайки този клас модели, П. Супес визира суровите данни, които изследователят получава като непосредствен резултат от проведените от него (емпирични) наблюдения. В този клас модели суровите, реални данни се представят в един непълен, но добре подреден, организиран и в известен смисъл идеализиран вид (Hambleton & Jones, 1993; Frigg & Hartmann, 2006). Нещо повече, „…моделите винаги предлагат непълна репрезентация на тестовите данни, за които са предназначени; по този начин, с достатъчно количество тестови данни, те могат […] да изглеждат непригодни” (Hambleton & Jones, 1993, стр. 254).

От друга страна, всеки модел се изгражда в рамките на определена психометрична теория, конкретизирайки в детайли взаимовръзките в мрежата от теоретични конструкти. В допълнение, всеки психометричен модел на данни се базира на определен набор от допускания, които по същество представляват описание на данните, за които е предназначен. В този смисъл тестовите модели са базирани на изследователския подход, който акцентира върху изграждането на такива модели, които съответстват на данните, а не върху издигането и проверката/ отхвърлянето на хипотези при установяване на липса на такова съответствие. Този подход е формулиран по изразителен начин от френския статистик Жан-Пол Бензекри като „втори принцип” при разработването на модели: „Моделът трябва да съответства на данните, а не обратно" (Greenacre, 1984, стр. 10).

Поради тези особености един от фундаменталните въпроси при прилагането на моделите на данни е за връзката, за съответствието, за „съвместимостта” между даден модел, по-точно между неговите допускания, и емпиричните данни, който може да бъде формулиран като въпрос за адекватността на модела (*model-data fit*). В този смисъл един теоретичен модел (в частност моделите на CTT и IRT), следва да се разглежда като приложим, ако има съответствие между допусканията на този модел и съответните характеристика на емпиричните данни. Р. Хамбълтън и Р. Джоунс, разглеждайки проблема за приложимостта на тестовите модели във връзка с техните несъвършенства, отбелязват, че „правилният” въпрос е не дали един модел е правилен или неправилен, а „…дали даден модел съответства на

данните достатъчно добре, за да бъде полезен при провеждането на измервателния процес" (Hambleton & Jones, 1993, стр. 254).

По-нататък, всеки модел има своя област на приложимост – онзи фрагмент от действителността, който моделът описва и спрямо който той е валиден. При моделите на CTT и IRT областта на приложимост следва да се разглежда като съвкупността от всички латентни променливи, чиито разпределения имат характеристики, съответстващи на допусканията на модела.

Необходимостта от съгласуване, от търсене на съответствие между теоретичния модел и наличните емпирични данни, се подчертава от много специалисти в областта на психологическите измервания (Lord, 1980; Crocker & Algina, 1986; Hambleton et al, 1991; Fan, 1998; Weiner et al, 2003). Загрижеността на авторите е свързана преди всичко с „твърдите", нееластични модели на IRT, които по своята същност имат огромен потенциал за решаване на различни изследователски и практически задачи. Техните предимства обаче могат да бъдат постигнати само тогава, когато „…съответствието между модела и тестовите данни е задоволително" (Hambleton et al, 1991, стр. 53).

Тестовите модели, особено тези в рамките на Теорията за отговор на тестов въпрос, са предмет на стотици публикации в психометричната литература. Авторите се фокусират или към техните теоретични аспекти, или представят приложенията им в най-различни емпирични контексти. Същевременно при много от приложенията на тестовите модели съответствието между даден модел и емпиричните данни, както и последиците от евентуалните несъответствия, не са изследвани адекватно или изобщо не са обект на изследователското внимание. К. Фан справедливо отбелязва „емпиричното безмълвие" в тази насока, което изглежда странно (Fan, 1998, стр. 359). Ъ. Уайнър обръща внимание на това, че много автори на статии, публикувани в реферирани списания, "…пренебрегват детайлното изследване на данните. Това води до лошокачествени ментални модели на феномените и предполага задължителна оценка на това дали данните се съгласуват с допусканията на параметричните тестове" (Weiner et al, 2003, стр. 36). Поради това сведенията за приложимостта за един или друг тестов модел в различни емпирични контексти са повече от оскъдни и поради това са необходими нови и

нови изследвания в тази насока (Kingston et al, 1985; Nandakumar, Yu, Li & Stout, 1998; Fan, 1998; Leeson & Fletcher, 2003; Hernandez, 2009).

CTT и IRT могат да бъдат разглеждани като функционално еквивалентни системи за измерване. Съответствието между тях се дължи на това, че имат една и съща област на приложимост, ориентирани са към един и същи сегмент от действителността, стъпват на една и съща емпирична база от данни и са предназначени за постигане на една и съща цел - скалиране и оценка на латентни черти. Р. Амарнани резюмира тази функционална еквивалентност в заглавието на една своя статия по следния начин: ”две теории, една тета” (Amarnani, 2009). По силата на функционалното им съответствие между двете теории съществуват редица сходства: споделят (някои) паралелни теоретични конструкти, процедури за разработване на психометрични инструменти и др. Най-важният паралел между двете системи е този, че при успоредното им прилагане на всеки индивид могат да бъдат приписани две различаващи се, но функционално еквивалентни числови стойности, отразяващи равнището на неговата компетентност.

Същевременно CTT и IRT са самостоятелни, независими, алтернативни и в този смисъл конкуриращи се теории, защото всяка от двете теоретични рамки предлага собствен набор от концептуални подходи, методи и техники за постигане на тази цел. Поради това е необходимо да се изследва не само съответствието между допусканията в техните модели и емпиричните данни, но и да се анализират в сравнителен план както техните „междинни” конструкти, свързани с описанието на психометричния инструмент на ниво тестов въпрос, субтест или цялостен тест, така и „финалните” им конструкти, отразяващи измерваната компетентност.

Както беше отбелязано, в специализираната литература се наблюдава дефицит и на преки сравнителни анализи на двете теории (виж Hambleton & Jones, 1993; Fan, 1998; Bechger, Maris, Verstralen & Beguin, 2003). В най-общ план психометричната общност е обединена около разбирането за безспорното превъзходство на IRT над CTT във всички аспекти на психологическите измервания (Hulin, Drasgow & Parsons, 1983; Hambleton & Swaminathan, 1985; Hambleton et al, 1991; Embretson & Reise, 2000; Baker, 2001). Р. Амарнани прави образно съпоставяне, сравнявайки двете теории с два фотоапарата, които правят различни снимки на едно и също нещо. CTT е

стар модел с малко пиксели, който прави сносни снимки, но с малък разход на енергия. IRT е съвършено нов модел апарат, който прави ярки, живи и контрастни снимки, но с голям разход на енергия (Amarnani, 2009). Поради несъмнените предимства на IRT, отбелязва К. Фан, би следвало да се очакват „значителни разлики между базираните на IRT и CTT статистики на въпросите и индивидите" (Fan, 1998, стр. 358). Авторите отбелязват недостатъчно добре проучените взаимоотношения между теоретичните конструкти в двете психометрични рамки, очаквайки в най-общ план определена монотонна връзка между посочените статистики (Lord, 1980; Crocker & Algina, 1986). Друг аспект на взаимоотношенията между CTT и IRT е проблемът за инвариантността (независимостта) на тестовите и личностовите статистики от извадките, в които са получени. Инвариантността се приема за едно от най-съществените предимства на IRT, което я отдалечава от CTT, в която съответните статистики са вариативни, и доближава измерването в нейните рамки до физическите измервания. Възниква въпросът доколко инвариантно е действителното поведение на тези статистики, получени в двете теоретични рамки

Беше отбелязано, че двете теории имат своите силни страни и своите ограничения, поради което всеки специалист в областта на психологическите изследвания следва ясно да заяви своето предпочитание и да работи в едната или в другата теоретична рамка. Разбира се, в своя избор той не е напълно свободен: решението трябва да бъде взето не "чрез основания *a priori*", а въз основа на конкретни емпирични доказателства за тяхната приложимост (Lord, 1980, стр. 14). Поради това адекватността на един или друг модел може да бъде оценена единствено във връзка с конкретна база от емпирични данни (Hambleton & Jones, 1993).

3. Общи наблюдения

Конкретните задачи, които си поставяме с този обзор, е да представим състоянието на изследванията върху приложимостта на CTT и IRT, главно върху съответствието между основните допускания за нормалност на разпределенията на латентните променливи, едномерност на латентното пространство и локална независимост на отговорите, и профила на данните в

областта на скалирането на когнитивни способности (постижения), както и някои други характеристики на двете теории. Трябва веднага да се отбележи, че такива проучвания са твърде оскъдни и, доколкото ни е известно, няма нито едно, в което тази проблематика да е обхвана в нейната пълнота и дълбочина. Може да се каже, че по-голяма част от резултатите по тази тема, които се отнасят предимно до IRT, са страничен продукт в значително по-обширната специализирана литература, посветена на параметричните статистики и на тяхната устойчивост. В допълнение, приложението на тестовите теории в много области извън психологическите измервания на постижения поставя различни от поставените, но не по-малко интересни въпроси. Поради това в обзора са представени и някои изследвания на приложимостта на тези теории извън заявената предметна област.

Скромният обем на изследванията на приложимостта, особено с реални данни, се подчертава от редица изследователи. В едно от най-често цитираните сравнителни изследвания на двете теории, представено по-нататък в текста, авторът К. Фан няколкократно подчертава „липсата на емпирично знание" относно поведението на различни статистики, базирани на двете теории (Fan, 1998, стр. 357). Литературното проучване, направено от автора, показва, че към онзи момент е направено само едно такова сравнително изследване между CTT и един от моделите на IRT, и то върху ограничен емпиричен материал.

Част от емпиричните изследвания са посветени на възможностите на двата метода за създаване на еквивалентни тестове, като резултатите са твърде пъстри и неоднозначни. Недостатъчни и несистемни са изследванията, които да третират въпросите за инвариантността на различните статистики в рамките на двете теории. Резултатите от ограничения брой анализи сочат по-скоро липса на инвариантност на статистиките както при CTT, така и при IRT. Изразявайки учудването си от оскъдния брой емпирични изследвания на тази тема, К. Фан допуска, че причината е в това, че „…превъзходството на IRT над CTT […] се възприема като даденост от психометричното общество", което не намира за необходимо да провежда внимателни проучвания (Fan, 1998, стр. 360).

Един от важните въпроси в тази област е за формата на разпределенията на изследваните психологически променливи. Въпреки

повсеместното използване на нормалната Гаусова крива за тяхното моделиране, емпиричните изследвания, посветени на формата на тези разпределения, получени в различни области, включително и в областта на социалните и хуманитарните науки, както и в образователните измервания, са сравнително малко. Т. Мичери говори за „потресаваща липса на такива данни от тестове за постижения и психометрични измервания" (Micceri, 1989, стр. 156). Както беше отбелязано, по-интензивни са изследванията на устойчивостта (нечувствителността) на някои параметрични статистики при нарушаване на техните основни допускания, включително и за нормалност на изходните разпределения, но те рядко са свързани конкретно с двете психометрични теории.

Въпреки че редица изследователи апелират към необходимостта от проверка на допускането за нормалност при всяко конкретно изследване (Lord, 1980; Kingston et al; 1985; Crocker & Algina, 1986; Breckler, 1990; Hambleton et al, 1991; Nandakumar, Yu, Li & Stout, 1998; Fan, 1998; Weiner et al, 2003; Leeson & Fletcher, 2003; Kline, 2005; Hernandez, 2009), то е сравнително рядко съблюдавано. В свое обширно изследване на приложението на структурното моделиране в психологията, С. Бреклер прави преглед на 72 статии в областта на личностовата и социалната психология, публикувани в 4 авторитетни американски психологически списания през 10-годишен (1977 – 1987) период. Анализирайки нарушенията на допускането за многомерна нормалност на разпределенията, авторът установява, че едва в 14 (19%) от статиите това изискване е посочено, и само в 7 (10%) от тях е обсъдено дали то е удовлетворено (Breckler, 1990). В останалите над 70% от публикациите авторите избягват този въпрос, най-често приемайки *a priori* изследваните променливи като нормално разпределени. Друг литературен преглед на статиите в 17 списания, направен от Х. Кеселман и сътрудници, показва, че авторите рядко верифицират статистическите допускания и използват модели, които не са устойчиви срещу нарушенията на тези допускания (Keselman et al, 1998, по Weiner et al, 2003).

Липсата на по-интензивни изследвания по поставения проблем би могла да се дължи на две основни причини: (1) липсата на достатъчно емпирични данни и/ или на достъп до налични данни (Micceri, 1989) и (2) убедеността в универсалния характер на нормалната крива, която апроксимира

разпределенията на много естествени феномени, включително и на психологическите променливи.

В много от публикациите се обръща внимание не само на нищожния обем на изследванията, посветени на проверката на едно или друго допускане, но и на обстоятелството, че по-голяма част от проведените изследвания са върху симулирани данни, основаващи се или на асимптотичната теория на екстремумите, или на изследвания по метода *Monte Carlo* на определени математически функции. Малко са тези, които, по думите на Т. Мичери „се осмеляват" да работят с данни от конкретни емпирични изследвания (ibid, стр. 158). Проблемът според него е в това, че характеристиките на математическите функции „се срещат рядко в реалните разпределения", получени в емпиричните изследвания (ibid, стр. 163-164). Такива симулирани изследвания, твърди авторът, „могат и да не представят реалните данни в каквато и да е разумна степен" (ibid, стр. 11). Проблемът, който поставя Т. Мичери, се отнася за външната валидност на симулираните изследвания, т.е. до каква степен изводите, направени въз основа на анализа на „изкуствени" данни, могат да се генерализират и най-вече да се пренесат и върху реалните данни. Авторът е на мнение, че някои оптимистични резултати от изследванията със симулирани данни следва да се приемат с известен скептицизъм поради това, че тези данни имат характеристики, различни от тези на реалните данни; параметричните статистики проявяват различни свойства при двата типа условия и, накрая, различни са и причините за проявите на тяхната (не)устойчивост.

Втората причина се корени в схващането за широкото разпространение на нормалната крива, което намира израз в често срещаните в различни публикации формулировки, че в типичния (или в конкретния) случай са налице достатъчно основания да се предполага, че случайните величини в психологията се подчиняват на нормално разпределение или поне не се отклоняват съществено от него. Един краен, но особено афористичен израз на това схващане е твърдението на Дж. Глас и К. Хопкинс: „Щастливо стечение на обстоятелствата е, че измерванията на много променливи във всички дисциплини имат разпределения, които са добра апроксимация на нормалното разпределение. Казано по друг начин, „Бог обича нормалната крива!" (Hopkins & Glass, 1978, стр. 95).

Макар и не с толкова висока чуваемост, има мнения на привърженици на противоположното становище, които поставят под съмнение идеите за широкото разпространение на нормалното разпределение в реалните емпирични данни, подхранвани от влиятелните разработки на Р. Фишер. „Може ли да бъде прието допускането за универсална нормалност?" пита риторично Е. Леман по повод идеите на Р. Фишер, обосновавайки съмнението си с тясната, несигурна емпирична база (наблюдения върху селскостопански култури), върху която последният гради методите си (Lehmann, 2008, стр. 117). Дж. Нунали споделя: „Строго погледнато, тестовите балове са много рядко нормално разпределени" (Nunnally, 1978, стр. 160). Авторът аргументира твърдението си с високата корелация между айтемите, която по необходимост трябва да присъства в един психологически инструмент и която води до разпределения с ексцес, по-нисък от нормалния.

Р. Гиъри разглежда парадигмата, установена от Р. Фишер като „предразсъдък в полза на нормалността". Според автора нормалното разпределение е особен случай, една от многото форми на разпределения, но не и универсална характеристика на променливите. Съвсем не на шега той настоява на всички съществуващи учебници по статистика, както и на бъдещите издания, да бъде изписано следното предупреждение: „Нормалността е мит; никога не е имало и никога няма да има нормално разпределение." (Geary, 1947, стр. 241).

В съвременната специализирана литература се забелязва тенденция на нарастващо недоверие към нормалното разпределение, която, обаче, „заобикаля" психолозите и психометриците (Micceri, 1989, стр. 156). По-важно е, разбира се, какви са емпиричните свидетелства в полза на едната или другата теза.

Отражението, което имат нарушенията на допусканията на различните модели върху надеждността и валидността на получените резултати, не е достатъчно изяснено поради малкия брой изследвания по този проблем. В допълнение, получените резултати и изградените на тяхна основа мнения и изводи на авторите са доста разнопосочни.

Според някои изследователи параметричните статистики са сравнително устойчиви на грешки от I и II род при леки опашки на разпределенията и слаба асиметричност. При по-силно изразена асиметричност t-тестът за независими

извадки (с приблизително еднакъв обем) и F могат да бъдат устойчиви срещу грешки от I род (Hsu & Feldt, 1969; Wilcox & Charlin, 1986; Micceri, 1989). Дж. Брадли обаче посочва, че едноизвадковият t, както и ANOVA със случайни ефекти могат да бъдат неустойчиви срещу грешки от I род при големи извадки при различни случаи на разпределения, отклоняващи се от нормалното (Bradley, 1980). Т. Мичери обобщава, че изследванията върху устойчивостта (нечувствителността) на параметричните статистики, направени през последните години, показват податливостта (от умерена до абсолютна неустойчивост) дори най-популярните сред тях на нарушения на изискванията за нормалност (Micceri, 1989). За неустойчивост на средната стойност и стандартното отклонение говорят резултатите на Д. Андрюс, М. Хил и У. Диксън и др. (Andrews et al, 1972; Hill & Dixon, 1982).

Има данни, според които отклоненията на разпределенията от нормалността в посока към асиметричност водят до съществено намаляване на надеждността на резултатите, особено при тестовете, ориентирани към норма (Brown, 1996). Според други литературни данни нарушенията на нормалността при факторния анализ водят до подценяване на факторните тегла и надценяване на броя на латентните фактори (Embretson & Reise, 2000).

Някои от авторите вземат решение „в полза" на удовлетвореността на едно или друго допускане в резултат на компромис, често при очевидно противоречие между изискванията на съответния модел и профила на реалните данни, или привеждат косвена аргументация. Така например Р. Харви обосновава решението си за едномерност на данните от теста O*NET, предназначен за измерване на абстрактни поведенчески характеристики на работното място, при наличието на свидетелства за тяхна многомерност, на предишни изследвания на Ф. Драсгоу и С. Парсънз (Drasgow & Parsons, 1983), според които моделите на IRT са устойчиви дори и към значителни нарушения на техните изисквания, включително и към отклонения от стриктната едномерност (Harvey, 2003). В представената по-долу статия на Н. Кингстън и колеги, авторите вземат две важни решения – за размерността на данните и за формата на характеристичната крива при теста GMAT, в противоречие или поне при липса на убедителни доказателства за съгласуваност със съответните допускания. Решенията са взети въз основа на стабилността на тестовата скала при от изравняването на тестовите балове, а също и чрез

реферираne към друг подобен тест, в сравнение с който случаите на съгласуваност при GMAT са много повече (Kingston, Leary & Wightman, 1985).

4. Изследвания върху данни от тестове за постижения

Ще започнем прегледа на публикациите по поставения проблем с две статии, в които се прави по-общ сравнителен анализ на особеностите, преимуществата и недостатъците на двете психометрични тестови теории.

В своя теоретична публикация Р. Хамбълтън и Р. Джоунс представят обширно съпоставяне на двете основни теории, обект и на настоящата разработка – Класическата тестова теория и Теорията за отговор на тестов въпрос (Hambleton & Jones, 1993). Подчертавайки генерално предимствата на „новата" теория, която се развива интензивно през последните 50-60 години, авторите не пропускат да обърнат внимание на това, че CTT също се развива и прилага успешно в множество тестови програми.

Дискутирайки разликата между понятията „тестова теория" и „тестов модел", авторите отбелязват, че бидейки по-абстрактна и съдържаща понятия от по-високо ниво, една тестова теория не следва да се оценява от гледна точка на нейната полезност. И обратно, като конкретизация на една или друга тестова теория, със специфични определения на теоретичните понятия и техните връзки, приложимостта на един теоретичният модел следва да се проверява. Оценката на годността на модела следва да се прави върху конкретна съвкупност от тестови данни, с добре обмислена система от емпирични методи.

Авторите характеризират моделите в рамките на CTT като „меки модели" поради обичайното и лесно постижимо съответствие на техните допускания и реалните тестови данни. Обратно, моделите на IRT са определени като „твърди" поради противоположните причини – далеч по-малката вероятност за такова съответствие.

Тестовите теории и съпътстващите ги модели са от изключителна важност за практиката на образователните и психологическите измервания, защото предлагат рамка за третирането на редица важни проблеми като грешката на измерване, типа на връзката между способностите и тестовите

въпроси и редица други проблеми, позволяващи конструирането на тестове с предварително зададени и желани характеристики.

Съпоставяйки двете теории и техните модели, авторите намират редица сходства, но и съществени различия между тях. Много от моделите на CTT са фокусирани върху тестовия бал, свързвайки тази статистика с действителния бал, докато IRT работи на по-ниско равнище, обвързвайки способността на индивида с отговора му на всеки конкретен тестов въпрос. Поради това статистиките на въпросите в новата теория са разположени на същата скала, на която се намира и способността на индивидите. Авторите разглеждат насочеността на IRT към отделните въпроси като нейно очевидно предимство, даващо на изследователя изключително голяма гъвкавост при определяне на характеристиките на тестовите резултати на една или друга популация или при конструирането на тестове, предназначени за дадена популация.

Една група принципни различия, които дават огромно теоретично предимство на IRT, са тези, че статистиките на въпросите и на теста като цяло са независими от извадките, въз основа на които са определени, а оценките на способностите – независими от теста, чрез който са определени. При CTT съответните статистики са зависими от извадките, което намалява тяхната полезност, освен ако извадките не се доближават по обем до генералните съвкупности, за които са предназначени съответните тестове. Личностовият параметър (действителният бал) също е зависим от трудността на използвания тест, освен ако тестовете не са паралелни, което е трудно постижимо. IRT разполага и с редица особености като характеристична крива на въпроса и на теста, тестова информационна функция и др., които представляват мощни средства за анализ на тестовите данни.

Наред с безспорните си предимства, новата психометрична теория има и някои недостатъци. Като такива авторите отбелязват сложността на моделите и проблемите, свързани с оценката на различните параметри. Особено важен е проблемът за приложимостта на нейните модели, тъй като все още не е ясно как този проблем може да бъде решен, особено що се отнася до размерността на тестовете. Това важи с особена сила за еднопараметричния модел, който изглежда най-лесно приложим, поради ограниченията, които се налагат от неговите допускания.

През последно време много от психометриците започват да предпочитат да работят в рамките на новата теория. Този смяна в акцентите на психометричното общество се дължи на ясното разбиране на слабостите на CTT и потенциалните предимства на IRT, които авторите резюмират по следния начин:

1. Независимост на параметрите на въпросите от извадките, въз основа на които са оценени.

2. Оценки на личностовия параметър, независими от трудността на теста.

3. Свързване на тестовите въпроси с равнищата на способност.

4. Не изискват стриктни паралелни тестове за оценка на надеждността

Към предимствата на CTT авторите отнасят:

1. По-малки по обем извадки, необходими за извършване на анализите.

2. По-лек математически апарат.

3. Концептуално прост и ясен модел за оценка на параметрите.

4. Не изисква задълбочени анализи на годността на модела за осигуряване на неговата приложимост спрямо конкретни данни.

В статия под образното наименование „Две теории, една тета" Р. Амарнани подчертава като основна особеност на CTT обстоятелството, че теорията разглежда теста като отделна единица, в която всички въпроси, независимо от характеристиките си, имат еднакъв принос при формирането на тестовия бал (Amarnani, 2009). Обратно, при IRT въпросите имат относителна тежест, такава, че за всяка тета (Θ) като оценка на съответната психична черта, съществува претеглен тестов бал, който й кореспондира. Поради това основната разлика в двете теории е в информацията, която се използва при формиране на тази оценка – по-груба и неточна при CTT и съответно по-прецизна при IRT.

Сред недостатъците на CTT, които авторът посочва, са тези, че наблюдаваният бал е просто оценка на действителния бал (Θ); индивидуалните резултати от различни тестове не са пряко съпоставими поради разликите в трудността на тестовете; всички норми при критерийно-ориентираните тестове са повлияни от нормативната извадка. Всички тези

недостатъци се преодоляват от IRT, която борави с вероятности, които са по-лесни за съпоставяне и обработване.

От друга страна, IRT борави с широкия спектър на трудностите на въпросите, всеки от които се характеризира с определена информационна функция, която е асоциирана с определено равнище на тета, което индивидът притежава, ако е отговорил правилно на съответния въпрос. Въз основа на отделните информационни функции на въпросите се определя информационната функция на теста, чрез която се определят най-вероятните (максимално правдоподобните) оценки на индивидуалните тета. Тези оценки са свързани с възможно най-ниските стандартни грешки на измерването, което улеснява интерпретацията на резултатите.

Представяйки трите най-използвани модела на IRT – 1, 2 и 3-параметричен, авторът поставя на дискусия основанията за избор на модел. Според него това са три допускания – за едномерност, за еднаква дискриминативна сила на въпросите и за възможността за налучкване на правилния отговор.

Авторът отбелязва все по-разширяващото се поле на приложения на IRT, което включва компютърното адаптивно тестиране (CAT), изпитите с висок залог (*high-stake exams*) както и при анализа на политомични данни. В заключение Р. Амарнани оптимистично посочва, че CTT и IRT са просто два психометрични метода за извличане на действителните балове от неясните, мъгляви ментални феномени. Принципът на Хайзенберг за неопределеността обаче не бива да стои като проклятие над бъдещето на психологическите измервания, защото те, особено IRT, показват, че може да се извлича психологическа информация с нарастваща информативна стойност.

Една от най-често цитираните публикации, представящи сравнителни изследвания на CTT и IRT, е тази на К. Фан (Fan, 1998). Изследването е фокусирано върху статистиките на айтемите и личностовите статистики в двете теории, по-точно върху взаимовръзките между сходните параметри/ индекси и доколко те са инварианти по отношение на различни извадки. В него, разбира са засегнати и други проблеми. Посочвайки теоретичните предимства на IRT и подчертавайки недостатъците на CTT, авторът прави разумното предположение, че поради този контраст следва да се очакват значителни различия между съответните статистики, изчислени в рамките на двете теории.

Подчертавайки недостига на емпирично знание по тези въпроси, авторът базира своите анализи на емпирични резултати от теста *Texas Assessment of Academic Skills* (TAAS), предназначен за ученици от 11. клас. Този инструмент, администриран от щатските власти, е критерийно-ориентирана тестова батерия, която се състои от три субтеста: четене (48 въпроса), математика (60 въпроса) и писане, който включва както обективни, така и въпроси със свободен отговор. Данните са събрани от над 193 000 ученици, явили се на тестов изпит.

Подобно на Р. Хамбълтън и Р. Джоунс (Hambleton & Jones, 1993), К. Фан разграничава понятията „теория" и „модел", по-скоро „модел от по-висок ред" и „модел от по-нисък ред", като първият е по-малко рестриктивен по отношение на своите изисквания отколкото втория. Въпреки това авторът започва емпиричната част на своето изследване, като не отделя почти никакво внимание на основните изисквания/ допускания на IRT, макар че изрично подчертава важността на тяхната проверка спрямо конкретните данни. К. Фан отбелязва, че разпределенията на тестовите балове не е нормално и се наблюдава ясно изразен таванен ефект, макар и да не привежда никакви конкретни данни. Без да посочва експлицитно метода за определяне на размерността (подразбира се, че е факторен анализ) и въз основа на собствените стойности на първите три фактора, без да посочва конкретен критерий, той приема наличието на един доминантен фактор за всеки от анализираните субтестове.

За да изследва съотношенията между статистиките в рамките на двата модела, авторът формира поредица от случайни извадки с обем 1 000 и. л., формирани на различни основания – 40 случайни извадки, 80 извадки, формирани по полов признак и 80 извадки от лица с ниски/ високи постижения от теста, всички извлечени от базата с данни.

Резултатите от изследването на съгласуваността между оценките на личностовия параметър (X_i по CTT и Θ_i по IRT, изчислени по 1-, 2- и 3-параметричния модел) показват изключително високи коефициенти на корелация, които за различните субтестове, извадки и модели на варират от 0.966 до 0.997. Съпоставянето на индексите/ параметрите на трудност на въпросите в рамките на двете теории води до идентични резултати –

коефициентите на корелация варира от 0.901до 0.990, в по-голямата си част над 0.980. Малко по-ниски от тези, но все така достатъчно високи са корелациите между дискриминативната сила на въпросите, оценена по двата метода (от 0.600 до и над 0.900). Корелационните коефициенти обаче силно варират в зависимост от типа на извадката, на субтеста или на модела на IRT. Авторът заключава, че дискриминационните индекси/ параметри проявяват тенденция да бъдат по-слабо съпоставими в сравнение с оценките на личностовия параметър и на трудността на въпросите.

Авторът проверява допускането за инвариантност на индексите/ параметрите на въпросите, изчислявайки тези статистики въз основа на различните извадки от и. л. (напр. мъже – жени, ниски - високи постижения и т. н.) Резултатите показват, че за инвариантност на статистиките може да се говори не само при IRT, но и при CTT. Така например средните корелации на индексите на трудност по CTT са в рамките на 0.945 – 0.993, а за съответния параметър по IRT – между 0.862 и 0.991. Отново малко по-ниски са средните корелационните коефициенти при съпоставяне на индексите/ параметрите на дискриминативна сила на въпросите, като при едни от извадките, в съчетание с 3-параметричен модел, средната корелация на съответните параметри е едва 0.020 (p = 0.089). Авторът заключава, че ако има някаква тенденция, тя е в това, че индексите на трудност по CTT са малко повече инвариантни, при почти всички условия, от съответните параметри по IRT.

К. Фан резюмира резултатите от направените от него съпоставителни изследвания, като обобщава, че те не водят до дискредитиране на Класическата тестова теория от гледна точка на приписваните й слабости, най-вече на нейната негодност да осигурява инвариантни статистики. Обратно, те не подкрепят IRT в нейното мнимо превъзходство по отношение на същата особеност. Този аргумент в полза на IRT е породен поради вакуума, създаден от липсата на емпирични доказателства. В психологическите измервания теориите са важни, заключава авторът, но техните достойнства трябва да бъдат доказани чрез строги, детайлни емпирични изследвания.

Може би най-близко до обсъжданата тематика е изследването, направено от Н. Кингстън и неговите колеги от ETS (Kingston, Leary & Wightman, 1985). Неговата основна цел, заявена от авторите, е да се направи изследване на приемливостта на IRT върху теста *Graduate Management*

Admission Test (GMAT), разработен и администриран от ETS. Авторите дебело подчертават, че необходима предпоставка за използването на IRT в която и да е тестова програма е приемливостта (*feasibility*) на нейните модели.

За проверка на приемливостта на IRT авторите прилагат два допълващи се подхода: (1) да се направи оценка на съответствието между допусканията на конкретен модел на IRT (3-параметричен, логистичен) и данните и (2) да се направи оценка на степента, в която нарушенията на тези допускания могат да възпрепятстват неговото използване или, напротив, въпреки нарушенията как приложението на този модел би могло да подчертае, дори да подсили някои от важните особености на GMAT. Една от тези особености е стабилността на тестовата скала (*score scale*), постигната чрез процедурите на изравняване на тестовите резултати от различни варианти на теста, използвани в различни тестови сесии.

GMAT е тестова батерия, резултатите от която се използват като част от приемните процедури на много университети в САЩ, Канада и Европа за прием на студенти в магистърски програми по бизнес, счетоводство, финанси, управление, по-специално за програмите от типа MBA. Тестът се състои от два субтеста – вербален и количествен, като освен субтестовите балове се изчислява и общ бал. Тези три резултата се извличат от 6 тестови секции с фиксирано време за работа. Във вербалния субтест се включват следните 3 секции: разбиране при четене (25 въпр.), редактиране на изречения (25) и анализ на ситуации (20); в количествения - две секции с текстови задачи за решаване на проблеми (30 + 20) и една секция за боравене с данни (30). Общия брой на въпросите е 150[1]. Предназначението на теста, както и неговото съдържание и структура, го отнасят към категорията на тестовете за оценка на склонността към обучение, при които се експлоатира предиктивната валидност на съответния измервателен инструмент. Това го сродява с Теста по общообразователна подготовка, който, обаче, по своето съдържание и начин на конструиране е типичен тест за постижения.

Авторите подлагат на проверка две допускания: (1) за едномерност на всеки от двата субтеста (вербален и количествен) и (2) за формата характеристичната крива на въпросите: логистична, която може да бъде

[1] Структурата на тестовата батерия, описана от авторите, е валидна към момента на събиране и обработване на данните. Към днешна дата неговата структура, както и броят на въпросите в отделните секции на GMAT, са различни.

описана с три параметъра (т.е. за приложимост на 3-параметричния логистичен модел).

Авторите боравят с реални данни, получени в процеса на провеждане на тестовите сесии и на изравняване на тестовите резултати, получени при администриране на различни негови варианти. Обект на изследване са 2 тестови варианта и 6 извадки от изпитани лица: две случайни извадки; две, дефинирани по полов признак (мъже и жени) и други две, определени по възрастов признак (една от млади хора на възраст 21 – 23 год. и една – от по-възрастни хора на 29+ год.) За всяка от извадките са анализирани 2 100 – 2 600 отговори на изпитаните на всеки тип въпрос, от всеки вариант на теста.

Направени са оценки на параметрите на въпросите отделно за вербалния и за количествения субтест. Получените оценки от двата субтеста са представени в отделни метрични пространства.

Представяйки кратък преглед на възможните методи за определяне на годността на 3-параметричния модел, авторите отбелязват, че въпреки усилията на изследователите, досега не е разработен задоволителен тест за оценка на съответствието между този модел и реалните тестови данни. Поради това, според авторите, „оценката на годността на модела е все още повече изкуство, отколкото наука” (ibid, стр. 15).

За постигане на поставените цели Н. Кингстън и неговите колеги използват последователно 6 различни методологически подхода, главно за оценка на размерността на субтестовите данни, на допускането за локална независимост и за формата на характеристичната крива на въпросите.

За оценка размерността на данните авторите правят повторен анализ на предходен изследователски факторен анализ на данни от същия тест. За да се подсигурят срещу евентуални негативни резултати, авторите правят уговорката, че 3-параметричният модел изисква едномерност, но това не предполага непременно линейна връзка между латентната променлива и тестовите въпроси. Макар че нелинейният факторен анализ съществува като теоретична концепция, на практика се работи с добре познатия линеен факторен анализ. Поради това резултатите от него могат да хвърлят светлина върху поставения проблем, но не могат да дадат дефинитивен отговор на въпроса за размерността.

Предходният факторен анализ по метода на главните оси е направен през 1981 год. от С. Суинтън и Д. Пауърс, върху данни от три варианта на GMAT. Изследователите установяват 6-факторна структура при всеки вариант на теста, като 5 от факторите имат еднаква интерпретация при всеки от тях. За целите на изследването Н. Кингстън и неговите колеги анализират повторно някои от получените резултати, като и за двата субтеста (вербален и количествен) отново получават многомерни латентни структури. При вербалния субтест – 4 факторна структура (два главни и два второстепенни фактора), а при количествения – 2-факторна структура. Появява се и слаб 7-ми фактор, който включва въпроси и от двата субтеста.

Вторият подхода е да се изследват взаимовръзките (корелациите) между въпросите в рамките на всяка от 6-те секции на теста. Резултатите от корелационните анализи предоставят друг вид допълваща информация за неговата размерност. Резултатите сочат, че корелациите между въпросите от двете секции на количествения субтест са относително високи, което е свидетелство за измерване на едни и същи характеристики (стойностите на r са 0.31 – 0.98). При трите секции на вербалния субтест обаче корелационните коефициенти са относително ниски (0.23 – 0.82, с преобладаващ дял на стойности под 0.50), което е свидетелство за това, че с тези секции се измерват различни характеристики.

Следващият анализ е фокусиран върху формата на характеристичната крива на въпросите и е осъществен чрез изследване на регресията на въпросите върху способността (*item-ability regression*). Това е графичен метод за съпоставяне на регресията на наблюдавания дял на правилните отговори на даден въпрос върху оценката на Θ (емпирична регресия) с характеристичната функция на съответния въпрос, определена въз основа на оценките на параметрите (оценена регресия). За тази цел авторите разделят скалата на способностите Θ (със средна стойност 0.00 и стандартно отклонени 1.00) на 15 интервала с ширина 0.40), като наблюдават дела на правилните отговори във всеки интервал. Като цяло резултатите от съпоставянето на двете криви са негативни, поради което авторите приемат нормативна тактика, съпоставяйки своите резултати с тези от *Graduate Record Examinations* (GRE) *General test*, който съдържа същите субтестове. Сравнението е полза на GMAT, при който

при въпросите от вербалния субтест се наблюдава малко по-добро, а при тези от количествения субтест – много по-добро съгласуване между емпиричните и съответните теоретични криви. Авторите обясняват по-добрите резултати при GMAT с по-хомогенната популация от изпитани и правят извода, че въпреки многомерността на латентната структура на субтестовете, 3-параметричната логистична функция на въпросите апроксимира добре данните от GMAT.

Следващата аналитична процедура е основана на статистическия тест Q_1, разработен от У. Йен (Yen, 1984) конкретно за проверка на годността на 3-параметричния логистичен модел. Тестът е модифициран от авторите съобразно данните, с които боравят, но резултатите съдържат висок процент на такива стойности на тестовата статистика, асоциирани със съответната вероятност от допускане на грешка от I род, показващи липса на съгласие между 3-параметричния логистичен модел и тестовите данни.

За да определят до каква степен многомерните тестови данни, обработени с едномерен 3-параметричен модел, въздействат върху оценките на параметрите на въпросите, авторите съпоставят тези оценки, изчислени въз основа на (1) хомогенни и (2) нехомогенни групи от въпроси. Хомогенни са, например, въпросите от всяка отделна секция на вербалния субтест, а нехомогенни – въпросите от целия субтест. Тяхното очакване е между двете групи оценки да има съществена разлика, тъй като група от еднородни въпроси е по-близо до едномерността, отколкото група от хетерогенни въпроси. Резултатите показват, че единствено параметърът b (трудност) не се влияе съществено от (не)хомогенността на групата въпроси, въз основа на които е изчислен. Това не се отнася обаче до параметрите a (дискриминативна сила) и c (налучкване), които демонстрират различно поведение в зависимост от групата въпроси – корелационните коефициенти при първия от двата за различните секции на теста са между 0.82 и 0.98, а на втория – между 0.69 и 0.96. Това, разбира се, може да се оцени като индикация за наличие на многомерни структури в тестовите данни. Подобни са резултатите и при съпоставянето на параметрите на въпросите, оценени върху включените в изследването различни извадки. Наблюдават се както високи, така и относително ниски корелационни коефициенти (0.40 – 0.45).

В заключение авторите отбелязват, че допускането за едномерност на вербалния и количествения субтестове не е удовлетворено от данните – при

двата субтеста се наблюдават многомерни структури, съставени от по два главни фактора и вероятно няколко второстепенни. Въпреки това обстоятелство някои от анализите показват, че трипараметричната логистична крива апроксимира достатъчно добре емпиричната регресията на наблюдавания дял на правилните отговори на даден въпрос върху оценката на Θ. Други анализи обаче показват отклонения от тази форма при някои от секциите на теста.

Вземайки предвид направените анализи и получените резултати, авторите правят изненадващото обобщение, че изследването има позитивни резултати и че е дало доказателства за приложимостта на IRT върху GMAT. Според тях, избраният модел съответства адекватно на данните от теста, независимо от факта, че GMAT се разработва като хетерогенна тестова батерия и очевидното нарушаване на ключови допускания на теоретичния модел.

В своя публикация Р. Нандакумар представя изследване за оценка на годността на психометричния софтуер DIMTEST да разкрива едномерни латентни структури при дихотомични данни (Nandakumar, 1993). Съществото в таза публикация е, че авторът изследва качествата на психометричния алгоритъм върху реални, а не симулирани тестови данни. В неговата сърцевина стои статистическият тест за оценка на основната едномерност (*essential unidimensionality*) на латентното пространство на тестови данни, разработен от У. Стаут (Stout, 1987). Авторът базира изследванията си предимно на резултати от тестовата батерия за оценка на склонността към обучение *Armed Services Vocational Aptitude Batery* (ASVAB) използвана при кандидатстване във въоръжените сили и военните училища в САЩ; от теста по математика ACT *Mathematics usage* (*The American College Testing Program*), използван за прием в някои американски колежи, който включва въпроси по алгебра, геометрия и тригонометрия; от теста ACT *Science*, измерващ умения за разчитане на графики и за интерпретация на данни в таблици, графики и фигури; от тестове за разбиране при четене, литература и американска история за 11 клас, както и от тестове за автотехници. Обемите на (суб)тестовете са между 25 и 36 въпроса, а изпитаните лица за всеки (суб)тест – между 750 и 5 000 души. Анализите показват, че само една част от

изследваните (суб)тестове са едномерни. Авторът обяснява многомерните тестове с наличието на субгрупи от въпроси, реферирващи към различни съдържателни области, които формират отделни димeнсии. Интересно би било да се отбележи, че сред (суб)тестовете, изследвани от Р. Нандакумар, които имат съответствие в ТОП, само при теста по литература проверката на съответната хипотеза индикира основна едномерност, докато тези по история и математика е по-вероятно да са многомерни. Авторът резюмира резултатите от изследването, заключавайки, че нито един от тестовете не се характеризира със стриктна едномерност. При всеки от тях се наблюдават, освен една основна, и няколко второстепенни димeнсии, които влияят върху отделни групи от въпроси. Някои от второстепенните димeнсии също могат да имат съществено влияние върху резултатите, домогвайки се до статута на основна димeнсия. Авторът завършва с парадоксалното твърдение, че „размерността на дадена група от въпроси е континуум" – не може да се определи със сигурност дали конкретно латентно пространство е едномерно или многомерно; то може да бъде само апроксимирано (ibid, стр. 36-37).

Р. Нандакумар, Ф. Ю, Х. Ли и У. Стаут представят интересно изследване за оценка на размерността (едномерността) на политомични тестови данни (Nandakumar, Yu, Li & Stout, 1998). Основната цел на изследването е да се оцени ефективността на психометричния софтуер Poly-DIMTEST (PD), разработен от двама от авторите (Х. Ли и У. Стаут) и предназначен за оценка на едномерността (или нейното отсъствие) при политомични тестови данни. Моделите на този тип данни представляват разширение на моделите на бинарни (дихотомични) данни и предполагат класифициране на индивидите в няколко последователни (вместо в две) категории. В този смисъл това изследване може да се разглежда като продължение на представеното по-горе изследване на качествата на аналогичния софтуер, предназначен за дихотомично скорирани данни.

Авторите подлагат на проверка симулирани едномерни и двумерни политомични данни, получени от два типа въпроси – с еднакъв и с различен брой категории. Изследването е направено при следните експериментални условия: две извадки със съответно 500 и 1 000 и. л. и два теста с максимален бал съответно 52 и 32 точки. Изводът на авторите е, че независимо от типа на

данните, алгоритъмът на PD успява да потвърди предварително зададената едномерност или да я отхвърли, ако симулираните данни са двумерни.

С. Райс анализира два статистически метода за изследване на съгласуваността на моделите на IRT и тестовите данни (Reise, 1990). Авторът обръща внимание на това, че за оценка на съгласуваността на тестовите въпроси (за всички изпитани лица) и за оценка на съгласуваността на отговорите на дадено лице (за всички въпроси) с конкретен модел, следва да се използват различни методи. Във фокуса на анализа авторът поставя две статистики - χ^2 за оценка на съгласуваността на тестовите въпроси и метода на максималното правдоподобие – за оценка на съгласуваността на личностовия параметър, с 3-параметричния логистичен модел на IRT. За целта С. Райс използва 9 матрици с дихотомични (1/ 0) данни, симулирани в съответствие с този модел. За да направи съпоставка на тяхното поведение, авторът прилага паралелно двата индекса както за оценка на въпросите, така и за оценка на Θ. Резултатите сочат, че двата индекса водят до почти еднакви резултати – около 94% „правилни" решения при оценка на съгласуваността на въпросите и около 97% - при оценка на личностовия параметър.

Все пак като цяло индексът χ^2 проявява тенденция към надценяване на броя на въпросите и на лицата, които не се съгласуват със заложения 3-параметричен модел, поради което препоръката на автора е методът на максималното правдоподобие да бъде използван и за двете цели.

Р. Хернандез представя емпирично сравнително изследване на двойка съответни индекси/ параметри на въпросите в CTT и IRT – дискриминативна сила и трудност (Hernandez, 2009). Авторът отбелязва, че анализът на качествата на тестовите въпроси е критичен момент в процеса на конструиране на психологическите тестове, по-голяма част от които се разработват в рамките на CTT. Разглеждайки слабостите на тази теория, авторът поставя въпроса дали при разработване на нови инструменти изследователите не биха могли да се възползват от предимствата на IRT.

Данните, с които борави Р. Хернандез, са от теста *Quick-Mental Aptitude Test* (Q-MAT), специално разработен за целите на неговото изследване. Тестът е с обем от 40 въпроса, групирани с два субтеста – вербален и невербален. Надеждността на теста на субтестово и тестово ниво не е много висока - KR-20

има стойности за r_{verbal} = 0.39, $r_{nonverbal}$ = 0.69 и r_{total} = 0.71. В изследването вземат участие 400 колежани, но броят на валидните/ анализирани тестове е 229.

Р. Хернандез определя стойностите на p (трудност), D и r_{pb} (дискриминативна сила) на въпросите по КТТ, съответно на b (трудност) за 1-, 2- и 3-параметричния модел на IRT и a (дискриминативна сила) за 2- и 3-параметричния модел. За определяне на връзките между съответните едноименни индекси/ параметри, авторът използва Пиърсъновия коефициент на корелация.

Като цяло резултатите сочат статистически значима, висока корелация между трудността и дискриминативната сила на въпросите, оценени по двата метода. Така например при съпоставяне на p и b за вербалния субтест корелационните коефициенти за 1-, 2- и 3-параметричния модел са съответно 0.857, 0.896 и 0.902 (значими при $p < 0.01$). Подобни са стойностите и при невербалния субтест (0.820, 0.984 и 0.974, при същото ниво на значимост). При съпоставянето на дискриминативните индекси D и a обаче се наблюдават и някои неочаквани инверсии. Корелационните коефициенти за 2- и 3-параметричния модел при вербалния тест са съответно 0.0891 (значим при $p < 0.01$) и -0.197 (без статистическа значимост), а при невербалния – съответно 0.945 (значим при $p < 0.01$) и 0.373 (без статистическа значимост).

Авторът открива и висока, положителна връзка между трудността на въпросите, определена по СТТ и съответно по трите модела на IRT, изчислена чрез коефициента на детерминация R^2, като най-високата му стойност при вербалния тест се наблюдава при 3-параметричния модел (R^2 = 0.81), а при невербалния тест тя е още по-висока, но при 2-параметричния модел (R^2 = 0.96). Висока положителна корелация при двата субтеста се наблюдава и при съпоставяне на дискриминативната сила на въпросите D и a, като по-високи стойности R^2 приема при 2-параметричния модел.

Авторът заключава, че има достатъчно доказателства за наличие на връзка между индексите/ параметрите, определени в рамките на двете теории, като се наблюдават определени различия. При невербалния тест корелациите са по-високи, отколкото при вербалния, а по отношение на моделите на IRT, 2-параметричният се съгласува по-добре както с трудността, така и с дискриминативната сила, определени по СТТ. Поради това изборът на автора пада върху 2-параметричния модел, макар че според изследване, направено

от С. Нукхет, 3-параметричният модел се съгласува най-добре със съответните индекси от КТТ (Nukhet, 2002), а според К. Фан такава съгласуваност се наблюдава без разлика при трите модела (Fan, 1998). В заключение Р. Хернандез препоръчва самостоятелното или паралелно използване на двете теории като рамки за разработване на тестове, по-специално на СТТ в условията на липса на специализиран софтуер или на сравнително малки по обем извадки.

М. Виберг представя резултатите от изследване в една на пръв поглед необичайна сфера (Wiberg, 2004). Авторката съпоставя възможностите на СТТ и IRT при анализа на въпросите от теоретичната част на теста за придобиване на свидетелство за управление на МПС (шофьорска книжка) в Швеция. Целта на изследването е да се направи оценка на годността на 1-, 2- и 3-параметричния логистичен модел на IRT за приложение върху резултатите от теста, а след това параметрите на въпросите по избрания модел да се съпоставят със съответните индекси по СТТ. Тестът е критерийно-ориентиран, с обем от 65 въпроса с множествен избор, обособени в 5 секции. Данните са получени от 5 404 кандидати за свидетелство, явили се на изпит.

В рамките на СТТ са изчислени надеждността на всяка секция (чрез коефициента α на Кронбах), както и индексите p (трудност) и r_{pb} (дискриминативна сила) на всеки въпрос. За IRT за изчислени съответните параметри b (трудност), a (дискриминативна сила) и c (налучкване на правилния отговор), като за 1- и 2-параметричния модел последните два параметъра са фиксирани на стойности съответно a = 1.00 и c = 0.00.

За преценка на това кой от моделите на IRT е адекватен на данните от теста, авторката прилага следните групи от критерии: (1) Верифициране на допусканията на модела, в които се включват (а) едномерност на данните, (б) еднаквост на дискриминативната сила на въпросите и (в) възможност за отгатване на правилния отговор; (2) Очаквани особености на модела, включващи (а) инвариантност на оценките на Θ по отношение на трудността на въпросите и (б) инвариантност на параметрите на въпросите по отношение на извадката от и. л.; (3) Годността на модела да предскаже актуалните тестови резултати чрез съпоставяне на действителните и предвидените чрез модела разпределения на тестовите резултати. Както се вижда, нормалността на

разпределенията не е сред допусканията, които авторката възнамерява да подложи на проверка.

Представяйки данните от CTT, авторката анализира големините на индексите на трудност и дискриминативна сила, отбелязвайки, че техните стойности варират значително, което й дава основание да заключи, че тези индекси следва да бъдат включени в моделите на IRT, макар и да не посочва въз основа на какъв количествен критерий прави този извод. Интересно е, че съгласно данните, които изследва, авторката не открива връзка между трудността и дискриминативната сила на въпросите по CTT, макар че също не посочва някаква количествена мярка, подкрепяща това твърдение. Авторката установява и необходимост от използване на параметъра за налучкване на правилните отговори. Тя съпоставя процента на лицата, попадащи в 10% извадка от и. л. с най слаби резултати, които са отговорили правилно на 5-те най-трудни въпроси, с теоретичните стойности на случайното налучкване. Макар че само при два от въпросите наблюдаваните проценти са по-високи от теоретичните, авторката заключава, че ефектът на налучкването е налице.

За оценка на едномерността авторката използва коефициента α на Кронбах и въз основа на неговата висока стойност при теста ($\alpha = 0.82$) прави извода за висока вътрешна консистентност и следователно за наличието на едномерна латентна структура. Като втори метод М. Виберг прилага факторен анализ, без да конкретизира неговия вид. Получената 65-факторна структура с 18 фактора със собствени стойности над 1.00 авторката интерпретира като едномерна поради наличието на един различим първи фактор, въпреки че той обяснява едва 9.00% от дисперсията.

М. Виберг установява наличието на локална независимост по метод, за който не дава пряка информация, не привежда и никакви доказателства. Следващото допускане, което се дискутира в текста, е, че въпросите в теста могат да бъдат моделирани чрез определен вид характеристична крива. В търсене на доказателства за съответствие на данните с модела авторката представя и анализира графиките на характеристичните криви на всички въпроси, генерирани чрез съответния софтуер. Тя обаче не дава информация по кой от трите разглеждани модела на IRT са получени графиките, нито коя (обща) особеност на характеристичните криви е търсеното от нея

доказателство.

За да провери допускането за инвариантност на оценките на способностите Θ_i, авторката разделя въпросите на две групи съобразно тяхната трудност (лесни и трудни), без да посочва по кой от моделите на IRT са направени оценките на b и как са формирани групите. След това изчислява индивидуалните Θ_i по трите модела на IRT въз основа на формираните субтестове по трудност. Резултатите са представени само графично, като диаграми на разсейването, въз основа на които авторката заключава, че това допускане не е удовлетворено. По подобен начин, разделяйки и. л. на две групи (с ниски и високи способности), М. Виберг проверява допускането за инвариантност на параметрите на въпросите. Само въз основа на графичния анализ тя заключава, че по отношение на трудността се наблюдава известна инвариантност (по-стриктна при 1-параметричния модел, по-малко стриктна – при останалите два модела). Различни са резултатите обаче по отношение на параметрите a и c, чиито оценки са далеч по-зависими от изпитаните лица.

Съпоставяйки оценките на параметрите на въпросите по IRT и съответните им индекси по CTT, М. Виберг установява висока корелация между параметъра a (дискриминативна сила по IRT, 3-параметричен модел) и индекса r_{pb} (по CTT), с коефициент на корелация 0.753. Корелацията между параметъра b (трудност по IRT, 3-параметричен модел) и индекса p (по CTT) има още по-висока и, както би могло да се очаква, негативна корелация от −0.861.

В заключение М. Виберг стига до извода, че нито един от анализираните три модела на IRT не съответства напълно на тестовите данни. От друга страна, всеки един от тях е по-подходящ от останалите по някои от своите параметри. Авторката се колебае в крайното си решение, но все пак предпочитанията й са към 2- и 3-параметричния модел, особено към последния поради високите стойности на параметъра c, които говорят за очевидната склонност на кандидатите за свидетелство за управление на МПС към налучкване на правилния отговор. Що се отнася до алтернативата IRT или CTT, авторката смята, че двете теории са полезни в еднаква степен, тъй като носят ценна информация както за теста като инструмент за измерване, така и за изпитаните лица.

О. Адедоин и сътрудници представят сравнително изследване на статистиките на двата теоретични модела (Adedoyin et al, 2008). Отбелязвайки, че CTT и IRT са представители на две съвършено различни измервателни концепции, авторите констатират, че „…са малко емпиричните изследвания, които са посветени на сходствата и различията в оценките на параметрите, получени при използване на двете теоретички рамки” (ibid, стр. 83).

Представяйки основните конструкти в двете теории, авторите подчертават като основни техни характеристики нестабилността на индексите на трудност и дискриминативна сила по CTT, тяхната зависимост от съответната извадка, и инвариантността на съответните параметри по IRT. Точно тази съпоставка дава основание на авторите да говорят за превъзходство на „модерния метод” за анализ на тестовите въпроси над „класическия”. Емпиричните доказателства в подкрепа на този предпоставен извод обаче са „твърде оскъдни” (ibid, стр. 85).

За да попълнят тази празнина, авторите си поставят за задача да изследват инвариантността на един от параметрите - трудността на въпросите (а) при различни извадки и (б) при различни обеми на извадките. Изследователските хипотези са проверени чрез MANOVA с повторни измервания. Данните са от теста *Junior Secondary School Certificate in Mathematics* (JSSC) за завършване на гимназиална степен на образование и са извлечени от извадка с обем над 36 000 и. л. От тази обща извадка авторите формират 155 субизвадки с еднакви и различни обеми, по признаците „пол”, „образователен регион” и „ниво на способности”.

За съжаление, авторите не посочват дали са направили съответните проверки за съгласуваността на данните с основните допускания, не посочват и по кой модел на IRT (и защо е предпочетен) са направили оценка на параметрите на въпросите.

Резултатите от анализа показват, че при малко над ½ от извадките се наблюдава инвариантност на индекса на трудност по KTT, докато при останалите, формирани главно по пол и образователен регион, инвариантност не се наблюдава. Резултатите от проверката на зависимостта на същия индекс от обема на извадката са по-категорични, също в полза на предположението за неговата инвариантност, с няколко изключения при извадки, формирани на регионален признак.

При анализа на инвариантността на съответния параметър b, оценен в рамките на IRT, резултатите категорично подкрепят тази теоретично обоснована особеност. Трудността на въпросите не се влияе нито от вида, нито от обема на извадката.

В заключение, О. Адедоин и сътрудници поставят под съмнение възможностите на CTT да осигури инвариантност на индексите на трудност и поради това препоръчват използването на IRT

5. Изследвания върху данни от други източници

Без съмнение, основната, традиционна сфера на приложение на двете тестови теории са образователните измервания. Все по-често обаче психометричните подходи се прилагат и в широката област на психологическите изследвания, а през последно време – и в сферата на медицината и здравеопазването, там, където е необходимо разработването и прилагането на различни типове въпросници и диагностични инструменти. IRT се прилага за изследвания в областта на медицинското образование (Brodin, Fors & Laksov, 2010), при дългосрочни проучвания на здравето на лица в юношеска възраст (Edelen & Reeve, 2007), при изследвания на общото функциониране на пациенти с деменция (Mungas & Reed, 2000) и др.

По-специално внимание заслужава обширното изследване на Т. Мичери (Micceri, 1989), който анализира характеристиките на разпределенията на 440 емпирични извадки, получени в различни области на социалните и поведенческите науки. Изследването е отклик на нарастващия интерес към устойчивостта (*robustness*) на параметричните статистики в условия на нарушаване на техните изисквания. Основанията на автора са, че след като има достатъчно доказателства за това, че параметричните статистики се характеризират с различна степен на устойчивост (чувствителност) към нарушенията на изискването за нормалност, това „наивно допускане" следва да бъде проверено, за да се определи какви са характеристиките на действителните разпределения (ibid, стр. 156)

Разпределения, с които борави Т. Мичери, са формирани при следните типове измервания:

(1) Тестове за общи постижения/ способности: 231 разпределения, извлечени от 20 различни теста: *California achievement tests*, *Comprehensive assessment program*, CTBS, *Stanford reading tests*, *Scholastic aptitude tests* (SAT), *Graduate record examination* (GRE), *College board subject area aptitude tests*, *American college test*, *Performance assessment in reading*, тестове на ETS за начинаещи учители, както и тестове от учебници, разработени от учители и др., изпълнени от лица от 45 различни популации. Предметните области са също разнообразни – езикови умения, количествени умения и логика, природни и социални науки, умения за учене, граматика и пунктуация.

(2) Критерийно-ориентирани тестове: 35 разпределения на тестови резултати от *Florida state assessment program* за ученици в областта на математиката и комуникационните умения от 3-ти до 11 клас, както и от *Florida teacher certification examination*, предназначен за оценяване на учители в областите четене, писане, математика и професионално образование, изпълнени от лица от 13 различни популации.

(3) Психологически тестове: 125 разпределения, включващи 20 различни скали: *Minnesota multiphasic personality inventory scales* (MMPI); въпросници за интереси; за гняв, тревожност, любопитство, мъжественост/ женственост, удовлетвореност, полезност, локус на контрола и др., изпълнени от лица от 21 различни популации.

(4) Резултати от ипсативни измервания (за установяване на разлики между резултатите от пре- и пост-тестове): 49 разпределения от 5 теста, изпълнени от лица от 10 различни популации.

Обемите на извадките, от които са получени разпределенията на тестовите балове, са 190-450 и. л. (10.8%), 460-950 и. л. (19.8%), 1000-4999 и. л. (55.1%) и 5000-10893 и. л. (14.3%). Около 90% от разпределенията включват 460 или повече наблюдения, а около 70% - над 1000 наблюдения. Възрастовото разпределение на и. л. включва 30.5% ученици до 6-ти клас, 20% ученици от 7-ми до 9-ти клас, 18.4% ученици от 10-ти до 12-ти клас, 9% студенти от колежи и 22% възрасни.

Тестовите балове от отделните измервателни инструменти варират от 10 до 99 точки (83.3% от всички инструменти). 12.5% имат тестов бал, по-нисък от 10 точки, а 4.3% - по-висок от 99 точки.

Разпределенията, с които борави Т. Мичери, са получени, по необходимост „според възможността да бъдат набавени", т. е. не са случайни (ibid, стр. 158). Източниците на данни са най-разнообразни - 15 големи издателства на стандартизирани тестове, изследователският департамент на Университета на Южна Флорида, Министерството на образованието на щата Флорида, няколко училищни региона в същия щат, както и автори на статии в множество авторитетни американски списания в областта на поведенческите и социалните науки като *Applied Psychology*, *Journal of Personality*, *Applied Psychological Measurement*, *Journal of Experimental Education*, *Journal of Educational Psychology* и др. Очакванията на автора са, че това разнообразие от източници би осигурило (почти) всички типове данни, обикновено получавани в емпирични условия.

Авторът прилага следните две групи от мерки за оценка на нормалността на разпределенията:

(1) Три мерки на асиметрия - M/M интервали (Hill & Dixon, 1982), асиметрия и Q_2 на Р. Хог (Hogg, 1974).

(2) Две мерки на теглото на двата края („опашките") на разпределенията - Q и Q_1 на Р. Хог и C-съотношение на Д. Елашоф и Р. Елашоф (Elashoff & Elashoff, 1978).

Въз основа на разработените от него критерии за оценка на нормалността на разпределенията, Т. Мичери прави детайлен анализ на масива от данни, от който тук ще представим само по-важните резултати, които кореспондират с настоящата разработка. Авторът установява, че при едва 15.2% от всички разпределения двете опашки имат тегла, равни или приблизително равни на контролните тегла на опашките при Гаусовото разпределение. При тестовете за постижения делът на разпределенията в норма е малко по-голям (19.5%), но това не може да се каже за нито едно (0.0%) от критерийно-ориентираните тестови разпределения. При психологическите измервания делът на разпределенията в норма е малко по-нисък (13.6%), а при ипсативните - 10.2%.

Резултатите от анализа на симетричността на разпределенията сочат, че 28.4% от всички разпределения могат да бъдат класифицирани като относително симетрични, а 30.9% - като крайно асиметрични. По типове измервания относително симетрични се оказват 34.2 % от тестовете за

постижения, 0.0% от критерийно-ориентираните тестове, 16.2 % от психологическите тестове и 53.1 % от ипсативните измервания.

Комбинирайки оценките за теглата на опашките и на асиметрията, Т. Мичери установява, че едва 30 (6.82%) от общо 440 анализирани разпределения се характеризират със стойности едновременно по двата показателя, които са близки до тези на Гаусовото разпределение. Броят на съответните разпределения от тестове за постижения е 23 (9.96%) от общо 231 разпределения, от критерийно-ориентирани тестове – нито едно (0.0%) от 35, при психологическите тестове - 4 (3.2%) от 125, и при ипсативните измервания – 3 (6.12%) от общо 49 разпределения.

Авторът не използва ексцеса като класификационен признак, не прилага и разработените за тази цел статистическите тестове за нормалност, тъй като, според него, тестването би било безсмислено и само по случайност би довело до решение за нормалност. Въпреки това той изчислява стойностите на ексцеса за всички разпределения, при което те варират в интервала (-1.70; 37.37). При 87% от разпределенията се наблюдават екстремни стойности на ексцеса (над 3.00), в по-голямата си част негативни, съпроводени от също така екстремни стойности на асиметрия. Интересно е, че авторът установява наличието на много висока, позитивна корелация между мерките на асиметрия и ексцес (r = 0.78).

В заключение, Т. Мичери не открива някакъв устойчив модел на разпределение на данните. В съвкупността от извадки се наблюдава широк диапазон на изменение на теглата на опашките на разпределенията, на тяхната (а)симетричност и изпъкналост, разпределения с една, две или три моди. Авторът заключава, че дори и при тези типове данни, получени в областта на поведенческите и социалните науки в резултат на прилагането на психометрични инструменти, всеки с фиксиран скалов диапазон, екстремните стойности на асиметрия и ексцес са по-скоро правило, отколкото изключение. Твърде малка част от разпределенията, според него, са „дори сравнително близка апроксимация на Гаусовото” (Micceri, 1989, стр. 161).

Р. Харви провежда интересно изследване на приложимостта на бинарния модел на IRT в областта на трудовата и организационната психология, по-конкретно за измерване на конструктите от категорията „обща

трудова дейност" (*general work activity*) в контекста на анализа на работното място и професията (*job and occupational analysis*) (Harvey, 2003).

В своето изследване авторът използва извадки от готов емпиричен материал, съхраняван в две национални бази данни, кумулирани от резултатите от два въпросника за анализ на работното място и професията - *Common-Metric Questionnaire* (CMQ), състоящ се от 25 айтема, измерващи различни видове физическа активност и *Occupational Information Network* (O*NET), включващ от 42 оценъчни скали, измерващи абстрактни поведенчески характеристики на работното място. Авторът подлага на анализ 6 507 профила на професии, събрани в базата CMQ, и 6 625 профила на професии, събрани в базата O*NET. Въпреки че Р. Харви изследва някои аспекти на приложимостта на IRT върху тези данни, изискването за нормалност на разпределенията на променливите не е сред тях.

За разкриване на факторната структура на въпросниците Р. Харви използва метода анализ на главни фактори, с квадрата на коефициента на множествена корелация в диагонала на корелационната матрица. Методът за извличане на факторите е на главните оси, с последваща неортогонална йерархична ротация по метода на Harris-Kaiser. При определяне на оптималния брой на факторите на въпросника O*NET Р. Харви прилага scree-теста на Кетел, в резултат на което идентифицира ясна 3-факторна структура. Съвкупно тази структура обяснява 91% от дисперсията, докато само първият фактор – 65%. По-нататък авторът селектира 18 (от първоначалните 42) айтема с най-високи факторни тегла по първия фактор, който е не само най-силен, но има и ясна интерпретация (дейности, свързани с междуличностните отношения). След като ги подлага повторно на факторен анализ по сходна на горната процедура, резултатите дават основание на автора да приеме еднофакторно решение. Авторът се обосновава не само с високата собствена стойност на първия фактор, но и с предходни изследвания, които показват голямата устойчивост на моделите на IRT срещу нарушенията на техните допускания, в частност – на допускането за едномерност. Според него, избраният модел се съгласува с допускането на IRT за „ефективна" едномерност поради обстоятелството, че първият фактор на новата латентната структура обяснява 88% от общата (споделена) дисперсия и 51% цялата дисперсия.

Заключението на автора е, че като цяло 3-параметричният модел на IRT е подходящ за приложение върху данните от двата въпросника. Това се отнася както за високите стойности на дискриминативния параметър a, за покриващите широк периметър от скалата стойности на трудността b, както и за ниските стойности на параметъра за налучкване c. Теоретичните характеристични криви също апроксимират достатъчно добре диаграмите на разсейване на съответните емпирични данни.

В своето изследване Р. Харви поставя интересният проблем за приложимостта на основното уравнение на 3-параметричния модел на IRT върху характеристиките на работата, идентифицирани в хода на анализа. Отбелязвайки, че традиционната сфера на приложение на IRT са когнитивните способности, при които параметрите b и Θ_i, както и съотношението между техните стойности, от които се определя вероятността от правилен отговор $P_i\left(\hat{\theta}\right)$, имат пряка и смислена интерпретация, това не е така в контекста на анализа на работното място и професията.

Авторът основателно допуска, че при някои димeнсии (конструкти) очакването, съгласно този модел, че лицата с високи стойности на Θ ще дадат положителен отговор на айтеми с по-ниски стойности на b, особено на по-„лесните", изглежда нереалистично. Например мениджърите от високите йерархични равнища на управление биха се съгласили, че вземат решения по стратегическото планиране на компанията („труден" айтем), но не и по въпросите за ежедневното разпределение на работата на служителите („лесен айтем"). Това, според автора, би предизвикало сериозни проблеми за точността на оценката на индивидуалните Θ_i.

Ситуацията, която авторът описва, може да бъде разгледана в перспективата на Кумбсовия модел QII „Данни единичен стимул" (Coombs, 1964). Възможно е данните при скали от този тип да бъдат от типа „отношение на близост", а не „отношение на подредба". Възможно е обаче проблемът да се отнася за формата на характеристичната крива (различна от „класическата" логистична крива) или до негативна различителна сила a на айтеми от този тип. Ако е вярно второто предположение, то въпросниците за изследване на обща трудова дейност биха могли да съдържат, най-общо, две групи айтеми: (1) отразяващи по-сложни работни задачи (с високи стойности на b и високи,

позитивни стойности на a), и (2) отразяващи по-прости работни задачи (с ниски стойности на b и високи, негативни стойности на a).

Проблеми с прилагането на класическата тестова теория, при която тестовият бал се формира като сумарна скала въз основа на скориране на отговорите като „правилни/ неправилни", авторът вижда при боравенето с отговорите от типа „не е приложимо/ не се отнася до мен" (*does not apply*) в случаите, в които даден въпрос не се отнася до определена категория лица, попадащи като цяло в целевата група. Обичайният начин за третиране на тези отговори, особено при използването на Ликертови скали, е да им се припише най-ниският бал, например нула при скала от 1 до 5. Проблемът възниква поради това, че при формиране на суровия бал класическата теория е чувствителна не към трудността на въпросите, а към броя на положителните отговори. Поради това могат да се получат парадоксални резултати, според които лица с действително по-високи стойности на Θ_i да получат по-ниски оценки (поради по-голям брой отговори от горния вид), а лица с действително по-ниски стойности на Θ_i - да получат по-високи оценки.

Решение на този проблем авторът вижда в използването на IRT и по-конкретно на 3-параметричния модел. Той обаче не посочва доказателства за своя избор (в сравнение с моделите с друг брой параметри), не дискутира и обстоятелството, че еднопараметричният модел на IRT също работи с броя на позитивните отговори. Не обяснява и как 3-параметричният модел би се справил с проблема със скалите, в които има отношения на близост в Кумбсовата теоретична рамка.

Изследвайки взаимовръзките между оценките на и. л. по IRT (Θ) и суровия тестов бал (X), Р. Харви наблюдава съществени разлики във формата на съответните разпределения – близко до нормалното при Θ и L-образно при X. При все това, авторът установява изключително висока рангова корелация между двете статистики (r = 0.97 при въпросника O*NET и r = 0.96 при CMQ). Макар че на равнище група се установява такава висока корелация, на индивидуално равнище се наблюдават значителни разлики между двете оценки, които достигат до 1 z-единица. Тези различия в оценките, според автора, биха имали значими последици за оценката на отделните индивиди.

Ф. Мангос и Дж. Джонстън представят интересно изследване на приложението на един от моделите на IRT (*Unfolding* IRT), базиран на Кумбсовата теория на данните (Coombs, 1964) за измерване на културни норми (Mangos & Johnston, 2008). Изследването е фокусирано върху психометричните качества на инструмента за измерване на културни норми и ценности *GlobeSmart Commander Self-Assessment Profile* (GS-SAP), и в частност - върху приложимостта на този психометричен модел върху получените данни. Въпросникът, който включва 32 айтема, разпределени в 6 субскали, е предназначен за оценка на влиянието на културните норми върху ценностите, намеренията, стремежите и поведението. В изследването участват 224 и. л., военни от 5 натовски страни, включително и от България.

Авторите открояват като важен един проблем, произтичащ от специфичната сфера на изследването, който е твърде сходен с този, който разглежда и Р. Харви – когато на измерване се подлагат психични характеристики, различни от способностите, прилагането на „стандартните" модели на IRT е проблематично. Проблемът се състои в това, че психологическият механизъм, който стои зад отговора на и. л. на даден айтем, съответства по-скоро на Кумбсовия модел „отношения на близост" между идеалната тока на индивида и точката на айтема на психологическия континуум на съответната черта, отколкото на модела „отношения на подредба". С други думи, вероятността от позитивен отговор не нараства монотонно, а има формата на нормална крива с център в идеалната точка на индивида.

Макар че изследването на приложимостта на конкретния психометричен модел е ясно заявено, авторите не поставят на обсъждане какви са неговите допускания, съответно не ги подлагат на проверка. При все това те установяват, че разпределението на оценките Θ на индивидуалните характеристики по отделните дименсии се подчинява на нормалното, макар и да не съобщават какъв метод за оценка са използвали и какви са конкретните резултати.

Цитирана литература

1. Adedoyin, O., Nenty, H., and Chilisa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review*, Vol. 3 (2), pp. 83-93.
2. Amarnani, R. (2009). Two тheories, one тheta: A rentle Introduction to Item pesponse theory as an alternative to Classical тest тheory. *The International Journal of Educational and Psychological Assessment*, Vol. 3, pp. 104-109.
3. Andrews, D. E, Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., Tukey, J. W. (1972). *Robust estimates of location survey and advances*. Princeton, NJ: Princeton University Press.
4. Baker, F. B. (2001). *The basics of Item response theory*. ERIC Clearinghouse on Assessment and Evaluation, 2-nd ed.
5. Bechger, T., Maris, G., Verstralen, H., Beguin, A. (2003). Using classical test theory in combination with item response theory. *Applied Psychological Measurement*, 27 (5), 319-334.
6. Bradley, J. W. (1980). Nonrobustness in z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, 16, pp.333-336.
7. Breckler, S. J. (1990). Application of covariance structure modelling in psychology: Cause for concern? *Psychological bulletin*, 107, 260-273
8. Brodin, U. Fors, U. and Laksov, K. (2010). The application of Item response theory on a teaching strategy profile questionnaire. *BMC Medical Education*, 10:14
9. Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.
10. Coombs, C. H. (1964). *A theory of data*. New York: John Wiley and sons, Inc.
11. Crocker, L., Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
12. Drasgow, F., & Parsons, C. (1983). Applications of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.
13. Edelen, M. O., Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16, pp.5-18.
14. Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
15. Fan, X. (1998). Item response theory and Classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, Vol. 58, No 3, pp. 357-381.

16. Frigg, R., Hartmann, S. (2006). *Models in science*. In: Edward N. Zalta (ed.) The Stanford Encyclopedia of Philosophy.
17. Geary, R. C. (1947). Testing for normality. *Biometrika*, 34, No. 3/ 4, pp. 209-242
18. Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
19. Hambleton, R. K., Jones, R. W. (1993). Comparison of Classical test theory and Item response theory and their applications to test development. *Educational measurement: Issues and practice*, 12 (3), pp. 535-556.
20. Hambleton, R. K., Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhoff.
21. Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of Item response theory*. Newbury Park, Ca.: Sage Publications, Inc.
22. Harvey, R. J. (2003). Applicability of binary IRT models to job analysis data. In Meade, A. (Chair), *Applications of IRT for measurement in organizations*. Symposium presented at the Annual conference of the Society for industrial and organizational psychology, Orlando.
23. Hernandez, R. (2009). Comparison of the item discrimination and item difficulty of the Quick-mental aptitude test using CTT and IRT methods. *The International Journal of Educational and Psychological Assessment*, Vol. 1, Issue 1, pp. 12-18.
24. Hill, M., Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics*, 38, pp. 377-396.
25. Hopkins, K. D., Glass, G. V. (1978). *Basic statistics for the behavioral sciences*. Englewood Cliffs, NJ: Prentice-Hall.
26. Hsu, T., Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance level of the F test. *American Educational Research Journal*, 6, pp. 15-527.
27. Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement*. Homewood, IL: Dow Jones-Irwin.
28. Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*, 68(3), pp. 350-386.
29. Kingston, N., Leary, L., Wightman, L. (1985). *An exploratory study of the applicability of Item response theory methods to the Graduate management admissions test* (RR-85-34). Princeton, NJ: Educational Testing Service.
30. Kline, T. J. (2005). *Psychological testing: a practical approach to design and evaluation*. Thousand Oaks, Ca.: Sage Publications, Inc.
31. Leeson, H., Fletcher, R. (2003*). An investigation of fit: Comparison of 1-, 2-, 3-parameter IRT models to project asTTle data*. Paper presented at the Joint NZARE/AARE Conference, Auckland.
Източник: http://www.aare.edu.au/03pap/lee03219.pdf

32. Lehmann, E. L. (2008). On the history and use of some standard statistical models. *Probability and Statistics: Essays in Honor of David A. Freedman*. Vol. 2 (2008) 114–126

33. Lord, F. M. (1980). *Applications of Item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

34. Mangos, P. M., Johnston, J. H. (2008). Applying Unfolding item response theory to enhance measurement of cultural norms. *Cultures and organizations: software of the mind. London*: McGraw-Hill.

35. Micceri, T. (1989). The Unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, Vol. 105. No.1, pp. 156-166.

36. Mungas, D., Reed, B. (2000). Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in medicine*, 19:1631-1644.

37. Nandakumar, R. (1993). Assessing essential unidimensionality of real data. *Applied Psychological Measurement*, Vol. 17, No. 1, pp. 29-38.

38. Nandakumar, R., Yu, F. Li, H., Stout, W. (1998) Assessing unidimensionality of polytomous data. *Applied Psychological Measurement*, Vol. 22, pp.99-115.

39. Nukhet, C. (2002) A study of Raven standard progressive matrices test's item measures under Classic and Item response models: An empirical comparison. Ankara University, *Journal of Faculty of educational science*, 35 (1-2), pp. 71-79.

40. Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

41. Reise, S. (1990). A comparison of Item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, Vol. 14. No. 2, pp. 127-137.

42. Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrica*, Vol. 52, pp. 589-617

43. Suppes, P. (1962). Models of data. In: E. Nagel, P. Suppes and A. Tarski (eds.). *Logic, methodology and philosophy of science: Proceedings of the 1960 International congress*. Stanford: Stanford University Press, pp. 252-261.

44. Weiner, I. B., Freedheim, D. K., Schinka, J. A., Velicer, W. F. (2003). *Handbook of Psychology: Research methods in psychology*. NJ: John Wiley and sons, Inc.

45. Wiberg, M. (2004). Classical test theory vs. Item response theory: An evaluation of the theory test in the Swedish driving-license test. EM, ISSN 1103-2685; 50. Umeå University, Faculty of Social Sciences, Statistics (Educational Measurement)

46. Wilcox, R. R., Charlin, V. L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics*, 11, pp. 263-274.

47. Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, pp. 125-145.