НБУ

НОВ БЪЛГАРСКИ УНИВЕРСИТЕТ

София 1618, ул. Монтевидео 21

Департамент „Когнитивна наука и психология”

ДОКТОРСКА ТЕЗА

на тема

Сравнителен анализ на приложимостта на две психометрични теории на тестовете

(върху данни от Теста за общообразователна подготовка)

Докторант: Любомир Георгиев Джалев, F58717

София, май 2013

Съдържание

Въведение

„Човекът е мяра на всички неща."
Протагор

Образованието е един от ключовите сектори в публичната сфера на модерните общества. Като специфичен вид обществена услуга, наред със културата, здравеопазването, социалното подпомагане и др., то изпълнява важни социални функции, като осъществява пренос на знания, умения, навици, нагласи и ценности от поколение към поколение. Фундаменталното значение на образованието е отразено във Всеобщата декларация за правата на човека[1], в която то, наред с правото на живот, свобода и сигурност на личността; на свобода на мисълта, съвестта и религията; на свобода на убежденията и на изразяването им, е провъзгласено за основно човешко право. В член 26 от този документ се казва: „Всеки човек има право на образование. […] Образованието трябва да бъде насочено към цялостното развитие на човешката личност и към засилване на уважението към правата на човека и основните свободи. То трябва да съдейства за разбирателството, търпимостта и приятелството между всички народи, расови или религиозни групи".

Образованието е не само индивидуално право, но и отговорност – в много национални законодателства обучението е задължително до навършване на определена възраст, обикновено асоциирана с някакво образователно равнище. Българската конституция също предвижда задължително и безплатно училищно обучение до 16-годишна възраст (съответстваща на 10-ти клас). Отговорностите, които модерната държава поема в областта на образованието, го превръщат в една от нейните най-институциализирани и централизирани структури. Образователната система осъществява възложените ѝ от обществото задачи чрез обучение, което е строго регламентирано, планирано и провеждано при относително известни и контролирани условия.

Често пъти се чуват гласове от образователната общност у нас, а и от медиите като изразители на публичното мнение, определящи образователната система като твърде консервативна. Такова мнение обаче трябва да се разглежда като амбивалентно. В него, от една страна, се съдържа признание за стабилността на системата, на нейната способност да се предпазва от социалните сътресения, да съхранява своя положителен опит и да пренася в бъдещето своите традиции, изграждани десетилетия. От друга страна, в него се крие упрек за (само)изолация, за затвореност и обърна-

[1] Universal Declaration of Human Rights. Adopted and proclaimed by General Assembly Resolution 217 A (III) of 10 December 1948.

тост към себе си. Образованието като че ли стои встрани от бързо развиващите се обществени процеси, бавно и трудно възприема новото и поради това изглежда закостеняло и дори старомодно.

Стремежът към устойчивост за сметка на динамичното развитие обаче поражда един въпрос, който се поставя пред образователната система все по настоятелно – въпросът за *качеството* на образователния продукт. Образователната система, а и семейството често са упреквани, че не успяват да се справят добре с мисията си на социализиращи институции. Независимо от това образователната система е разработила редица механизми за своевременен, текущ вътрешен контрол на процесите, протичащи в нея, най ефективният сред които е *оценяването* на знания и умения. Впрочем, образователната практика включва различни видове оценявания, които могат да бъдат категоризирани по множество признаци – вътрешно (базирано в класа или студентската група) и външно, синхронно и диахронно, по различни учебни дисциплини и образователни програми, с различен териториален обхват – локални и интернационални като програмите за оценяване TIMSS (*Trends in international mathematics and science study*), в която през 2011 г. са участвали 63 страни, PIRLS (*Progress in international reading literacy study*) с 49 страни участнички през същата година и PISA (*Programme for international student assessment*) с 65 страни участнички през 2009 г. Първите две програми, ориентирани към сферата на средното образование, се администрират от Международната асоциация за оценяване на образователните постижения (IEA), базирана в Амстердам, а третата – от Организацията за икономическо сътрудничество и развитие (OECD).

Университетското образование също е обект на особено внимание. В САЩ действат стотици институции и програми, чиято цел е поддържането на високи образователни стандарти, например *Baldridge national quality program*, която е насочена не само към сферата на висшето образование, но и към бизнеса. В Европейския съюз почти непосредствено след началото на Болонския процес е учредена Европейска асоциация за осигуряване на качеството във висшето образование (ENQA), която обединява независими обществени организации, неправителствени сдружения и правителствени агенции, включително и българската Национална агенция за оценяване и акредитация. ENQA работи за постигане на целите, прокламирани от Болонската декларация[2] за изграждане на „Европа на знанието", и най-вече за развитие на европейското сътрудничество за осигуряване на качеството чрез създаване на съпоставими критерии и методологии в рамките на Европейското пространство на висшето образование (EHEA).

Оценяването следва да се разглежда не просто като описващо постиженията, а като мощна движеща сила за промяна в образователната система, водеща до подоб-

[2] Подписана на 19 юни 1999 г. в Болоня от 28 европейски държави, включително и от България.

ряване на качеството и до по-високи стандарти на обучение (Wolf, Bixby, Glenn & Gardner, 1991). Усилията за постигане на образование, съответстващо на съвременните реалности, са обобщени в един от годишните доклади на ЮНЕСКО със символичното заглавие “Using assessment to improve the quality of education” (Kellaghan & Greaney, 2001). Разглеждайки различните подходи към този проблем, авторите отбелязват, че „през годините оценяването се превърна във важен ключ за подобряване на качеството на образованието. То е един от най-надеждните начини за установяване на проблемите, независимо дали са на системно равнище, на равнище училище или засягат отделния ученик или студент” (Kellaghan & Greaney, 2001, стр. 7). Авторите представят оценяването и по-конкретно оценяването на индивидуалните постижения на обучаваните в обща концептуална рамка заедно с три други базисни понятия - *качество*, *стандарти* и *отговорност*. Тука става дума за разпределяне и конкретизиране на отговорността за непостигане (или постигане) на действащите образователни стандарти на всички – преки и косвени - участници в образователния процес.

Тук следва да добавим още един важен аспект на отговорността, който би могъл да има сериозни практически последици – осигуряване на надеждността и валидността на информацията от оценяването. Ако възприемем като работна дефиницията за качество, която авторите предлагат: „…адекватност или уместност на обектите или процесите за постигане на целите, за които те са предназначени” (Kellaghan & Greaney, 2001, стр. 22), можем да преформулираме поставения проблем като *проблем за качеството на оценяване* на постиженията.

Днес факторите, отговорни за оценяването на резултатите от обучението, могат да прилагат различни методи в рамките на субективното и обективно оценяване, границата между които, между впрочем, не е толкова рязка, както обикновено се смята. Историческият развой на двата подхода, движен от стремежа за тяхното усъвършенстване, е довел до разгръщането на широка палитра от техни форми и методи за оценяване. Така например есето, един от най-популярните видове въпроси с конструктивен отговор, се използва в над 50 разновидности. Това разнообразие отразява разнообразието на изпитните ситуации и на изпитните цели, за постигането на които е подходяща съответната форма. Тук можем да си припомним афористичната мисъл на У. Попхем, един от теоретиците-класици на критерийно-ориентираното тестиране, според когото „...средствата, предназначени за една цел, рядко са годни за постигането на друга” (Popham, 1981, стр. 18).

Освен прагматични, поставеният проблем има и специфични изследователски измерения. Натрупаният практически опит намира отражение в изграждането на психометрични теории и модели с различна степен на обобщеност. Голяма част от тях, по-конкретно ориентираните към обективното оценяване тестови теории и съставящите ги модели, могат да бъдат класифицирани като „модели на данни”. Тяхна отличителна характеристика е, че включват в себе си определен набор от основни *допуска-*

ния, които могат да се възприемат като общо описание на данните, за които е предназначен съответния модел. От друга страна, при прилагането на даден модел неговите основни допускания следва да се разглеждат като *изисквания*, като необходими предпоставки за неговото използване. Поради това един от фундаменталните въпроси при моделите на данни е за връзката между съответния модел и емпиричните данни, който може да бъде формулиран и като въпрос за адекватността, за приложимостта на модела. Съгласно втория принцип за изграждане на модели на френския статистик Жан-Пол Бензекри, „…моделът трябва да пасва на данните, а не обратно" (цит. по Greenacre, 1984, стр. 10). Тази взаимна обусловеност поставя акцента в изследователската работа върху разработването на модели, подходящи за съответния тип данни. Това е акцентът и в емпиричната част на представената разработка, която е фокусирана върху анализ на приложимостта на две основни психометрични теории – Класическата тестова теория и Теорията за отговор на тестов въпрос, върху резултатите от Теста по общообразователна подготовка, който стои в основата на приемните процедури в Нов български университет.

I. Теоретична рамка на изследването

1. Психометрични тестови теории

Конструирането на теста е комплексен, продължителен и в някои свои фази – итеративен процес с изследователски и приложен характер, който включва етапите на планиране, разработване, апробиране и анализ на данните, както и оценяване на постиженията на изпитаните лица. Дейностите във всички етапи на конструирането и използването на теста се извършват в рамките на определена психометрична (тестова) теория. Напредъкът в развитието на психологическите измервания се изразява не само в интензивното развитие на теоретичните им основи, но и в екстензивното усъвършенстване на съществуващите и в разработване на нови психометрични теории и модели, чрез които измерването в поведенческите и социалните науки все повече се доближава до строгите изисквания на измерването в природните науки. В действителност (почти) всяка тестова теория съществува под формата на множество теоретични модели.

В настояще време е налице широка палитра от психометрични теории, които по правило са ориентирани към едни и същи данни, но се съревновават чрез различни подходи за тяхното моделиране.

- Класическа тестова теория (*Classical test theory*), фокусирана върху осигуряване на надеждността на резултатите от измерването на равнище цялостен тест.
- Теория на генерализацията (*Generalizability theory, G-theory*), представляваща развитие на Класическата теория, предназначена за осигуряване на надеждността и валидността на наблюденията чрез едновременно оценяване на множество източници на грешки в измерването (Cronbach et al., 1972; Cardinet et al., 1976; Shavelson & Webb, 1991; Brennan, 2001; Steyer, 2001).
- Теория на латентните състояния и черти (*Latent state-trait theory, LST theory*), представляваща развитие на Теорията на генерализацията, въвежда формални дефиниции на понятията „състояние” и „черта”, както и методи за тяхното разграничаване, отчита влиянието на факторите на ситуацията върху резултатите от измерването и разпростира този подход до анализ на отделни въпроси, базиран на нормалната огива (Steyer, Majcen, Schwenkmezger, Buchner, 1989; Steyer, Ferring & Schmitt, 1992; Eid, 1996; Steyer, Schmitt & Eid, 1999; Courvoisier, Eid, & Nussbeck, 2007)
- Теория за отговор на тестов въпрос (*Item response theory, IRT*), фокусирана върху анализа на резултатите на ниво тестов въпрос.
- Теория за отговор на група от въпроси (*Testlet response theory, TRT*), фокуси-

рана върху изследването на малки групи от еднородни въпроси (*testlets*, *content-dependent item sets*), които се разглеждат като основна структурна единица на теста. Теорията се базира изцяло на Байесовския подход за оценка на вероятностите, а параметрите се оценяват чрез използване на методите Монте Карло за Марковски вериги (Rosenbaum, 1988; Thissen, Steinberg, & Mooney, 1989; Wang et al., 2006; Downing & Haladyna, 2006; Wainer, Bradlow & Wang, 2007).

1.1. Класическа тестова теория

1.1.1. Основни идеи и понятия в CTT

Класическата тестова теория (CTT) е еманация на усилията на няколко поколения изследователи в областта на поведенческите и социалните науки за квантифициране на различни аспекти на индивидуалните различия. Нейните основи са поставени преди повече от 100 години от Чарлз Спиърмън в неговото знаменито изследване на интелигентността, представено в статията "General intelligence", objectively determined and measured" (Spearman, 1904). В тази статия, определяна като един от най-влиятелните текстове в областта на психометричните изследвания, Ч. Спиърмън лансира идеята, че резултатът от едно измерване може да се разглежда като отражение не само на измерваното свойство на даден обект, но и на особеностите (несъвършенствата) на използвания инструмент или процедура. Разглеждайки недостатъците в методологичните подходи на своите предшественици, изследвали интелигентността, Ч. Спиърмън отбелязва четири техни "смъртни гряха", между които и този, че "... нито един изследовател, изглежда, не взема под внимание друг голям източник на неточности, който неминуемо присъства във всяка работа, а именно *грешките на наблюдението*" (Spearman, 1904, стр. 223). След като експериментът бъде извършен и корелацията между променливите – изчислена, изследователят трябва да има предвид, че тя не представя точно количествените отношения между двете серии от сравнявани реални обекти, а само между двете серии от получени стойности. Резултатите от всяко изследване са повече или по-малко повлияни от различни обстоятелства, които експериментаторът не би могъл да предвиди или да контролира. Доказателство за това е фактът, че при повторение на едно наблюдение или експеримент могат да се получат резултати, различни от вече получените. Но изследователите обикновено пренебрегват това обстоятелство, констатира Ч. Спиърмън, водени от убеждението, че тези отклонения взаимно се компенсират: половината от тях действат в посока към увеличаване на „видимата" корелация, а другата половина - към нейното намаляване. Предполага се, че чрез взаимното уравновесяване на тези сили полученият резултат ще бъде много близо до „истинския" (Spearman, 1904).

Макар че в тази статия Ч. Спиърмън използва термина „надеждност" (*reliability*) само веднъж, въз основа на неговата евристична концепция за грешката на измерва-

нето възниква теорията, станала известна като „класическа" теория на надеждността. Съгласно тази теория, резултатът от едно психологическо измерване може да се третира като композитна стойност, съставена от два компонента: действителна стойност на измерваната характеристика и грешка, допусната при измерването. Ето защо Класическата тестова теория борави с три основни конструкта:

(1) Наблюдаван тестов бал (*obtained, observed, test score/ value*). Това е „видимият" резултат от измерването, стойността, която изследователят регистрира.

(2) Действителен, „истински" тестов бал (*true score/ value*). В концептуализирането на този конструкт могат да се очертаят два различни, дори взаимно изключващи се подхода. Единият от тях разглежда действителния бал като „чиста математическа абстракция", като понятие, което няма валенции към реалния свят (Стоименова, 2000, стр. 38). При математическото моделиране на зависимостите между действителния резултат и останалите компоненти не се цели да се определят параметрите на модела така, като че ли те действително съществуват в реалния свят. Противоположното схващане е, че понятието "действителен бал" има денотат и това е някаква "чиста", присъща на всеки индивид черта (Sax, 1989). В областта на измерванията в образованието и психологията действителният бал отразява "действителното" равнище на измервания признак – знания, умения, способности, атитюди или личностови черти, които характеризират индивида.

В Класическата теория действителният бал на индивида е относително устойчива, стабилна величина, която не се променя (поне за определен период от време) при многократно измерване с един и същи тест или с различни (еквивалентни, паралелни) форми на теста. На индивидуално равнище този компонент е константа, по-точно параметър с фиксирана, но неизвестна стойност. Ф. Лорд обаче прави разграничение между понятията за наблюдаван и действителен бал, от една страна, и способности – от друга (Lord, 1953). Разликата между тях е, че големините на първите две са зависими от конкретния тест, а на третото – че не е зависимо.

Макар че действителният тестов бал не може да бъде наблюдаван пряко, точно той стои във фокуса на Класическата теория, а и на изследователския интерес при провеждане на измервания за научни или практически цели. Белег за това е едно от синонимните й наименования – Теория на действителния бал (*True score theory*). В Класическата теория наблюдаваният бал служи за оценка на неизвестния действителен бал (Embretson & Reise, 2000)

(3) Грешка на измерването (*measurement error, error score*). Тази грешка е неизменен спътник на всяка процедура за измерване, независимо от предметната област, в която е извършена. Още по-характерни са грешките на измерването в областта на поведенческите и социалните науки. За нейното невидимо присъствие издайнически говорят повече или по-малко различните стойности, които можем да получим при многократно измерване на някаква характеристика даден индивид, дори и ако е проведено

с един и същи инструмент. Това означава, че всяко измерване се асоциира с определена грешка, която може да се разглежда като разлика между действителния и наблюдавания тестов бал или, с други думи, като отклонение на получения резултат от действителната стойност на измервания признак.

В Класическата теория наблюдаваният тестов бал се разглежда като сума от действителния тестов бал и грешката на измерването. Връзката между представените три конструкта се представя чрез следното *основно уравнение на* CTT (Crocker & Algina, 1986; Michell, 1999; Weiner et al., 2003; Kline, 2005; Boyle & Fisher, 2007)

$$X_i = \tau_i + \varepsilon_i \quad (1)$$

където:

X_i – наблюдаван тестов бал на *i*-тия индивид

τ_i *(tau)* – действителен тестов бал на същия индивид

ε_i – случайна грешка на измерването

Нека да разгледаме някои особености на този математически модел. Той представя действителния бал и грешката на измерването като адитивни компоненти на наблюдавания бал. Компонентите в дясната част на равенството са ненаблюдаеми пряко; единственият компонент, за който може да бъде получена количествена информация, е тестовият бал в лявата му част. Неговата стойност винаги е асоциирана с конкретен индивид, при конкретна процедура на оценяване, с конкретен инструмент.

Всяка процедура на измерване, която резултира в конкретна стойност, може да бъде представена като провеждане на два съвместни случайни експеримента: (1) подбор (*sampling*) на измерваем обект u (в рамките на психологическите измервания това е изследвано/ изпитвано лице), който принадлежи към дадена популация Γ_U и (2) подбор или регистриране на наблюдение o, което е елемент от множеството на възможни наблюдения Γ_O, каквито са например стойностите в границите на изменение на тестовия бал (Steyer, 2001). Множеството от възможни изходи на случайните експерименти се определя като Декартово произведение на множества:

$$\Gamma = \Gamma_U \times \Gamma_O \quad (2)$$

Елементите на Γ_O, регистрираните наблюдения, могат да имат различна форма - качествена, например „слаб” или "много добър", или количествена, т.е. да бъдат определена числова стойност, какъвто е индивидуалният тестов бал X_i. В областта на психологическите измервания резултатите обикновено се определят посредством конкретни за всеки инструмент правила, описващи как наблюденията трябва да се трансформират в числови стойности (тестови балове). Обикновено психологическите скали са сумарни (*sum scale*), т.е. получават се чрез сумиране на точките, приписани на все-

ки айтем, но могат да бъдат и по-сложни, каквито обикновено са при инструменти, в които айтемите се оценяват чрез скали от ликертов тип. Трябва да се отбележи, че Класическата теория не предписва определени правила за формиране на наблюдавания бал (Steyer, 2001), но независимо от това общоприетото практическо правило е тестовият бал да се образува като сумарна скала.

Нека да се върнем към основното уравнение (1), което, поставено в този контекст, може да бъде разгледано от две перспективи: на отделния индивид и на популацията, към която той принадлежи. Т. Клайн отбелязва, че това са два различни подхода за изграждане на теоретичния модел на наблюдавания и действителния бал: подход, базиран на идеята за многократно оценяване на един и същи индивид и подход, базиран на идеята за еднократно оценяване на множество индивиди (Kline, 2005). Тези два подхода, според автора, водят до един и същи математически модел и това „ускорява нещата драматично", давайки възможност за прилагане на практически по-осъществимия начин за събиране на данни съгласно втория подход (ibid., стр. 93).

При многократно измерване на даден индивид чрез един и същи тест или чрез паралелни (еквивалентни) тестове, наблюдаваният бал би приел различни стойности, което дава основание X_i да се разглежда като случайна променлива величина с известно вероятностно разпределение. Тогава действителният бал на индивидуално равнище може да се дефинира чрез математическото очакване (средната стойност) на наблюдавания бал $E(X_i)$ на съответния индивид. Средният (действителният) бал има ключово значение за Класическата теория, тъй като много от нейните основни допускания произтичат от тази дефиниция (Lord, 1980; Стоименова, 2000; Steyer, 2001).

$$E(X_i) = \tau_i \tag{3}$$

Разликата между действителния и наблюдавания бал се обяснява с грешката на измерване, която също може да приеме различни стойности при многократно измерване на същия индивид. Тогава ε също е случайна променлива величина, с неизвестно вероятностно разпределение (Lord, 1980).

$$X_i - \tau_i = \varepsilon_i \tag{4}$$

Грешката на измерването в Класическата теория се разглежда като статистическа случайна грешка (*random error*). Този тип грешки съпътстват неизменно всяка измервателна процедура и са резултат от непредсказуеми флуктуации в някой от елементите на измерването (несъвършенство на измервателния инструмент, влияние на тестовата ситуация върху индивида, неговото здравословно състояние или способност за концентрация, както и субективни грешки, свързани с некоректно прилагане на измервателните процедури, неправилно отчитане на получените резултати и др.) С други

думи, възникването на случайните грешки се дължи на въздействието на комплекс от случайни (неорганизирани) фактори, които измерващият субект не може да контролира. Наличието на грешка може да доведе до надценяване на действителния бал (тогава нейната стойност е положителна) или до неговото подценяване (грешката е с отрицателна стойност).

Математическото очакване на случайната грешка на индивидуално равнище е нулево, тъй като от предходните уравнения следва, че:

$$E(\varepsilon \mid \tau) = E(X - \tau \mid \tau) = E(X \mid \tau) - E(\tau \mid \tau) = \tau - \tau = 0 \qquad (5)$$

Нулевата стойност на математическо очакване означава, че в Класическата теория грешката на измерването се третира като несистемен компонент, следователно средният наблюдаван бал на едно лице $E(X_i)$ е неизместена оценка на неговия действителен бал τ_i (Lord, 1980; Стоименова, 2000).

Подобно на всеки друг обобщаващ теоретичен модел, Класическата теория не се интересува от индивидуалните случаи, а от "поведението" на компонентите на основното уравнение и отношенията между тях при големи групи от индивиди - генерални съвкупности (популации) или субпопулации.

Разгледан от тази перспектива, действителният бал τ е случайна променлива величина с неизвестно разпределение. Случайната грешка е също променлива величина с неизвестно разпределение, което се формира от стойностите й при отделните индивиди. Голяма част от допусканията в СТТ са формулирани от тази перспектива (Lord & Novick, 1974).

Действителният бал като случайна променлива се формулира като условно очакване на наблюдавания бал X_i при дадена променлива U:

$$\tau_i = E\left(X_i \middle| U\right) \qquad (6)$$

Тогава индивидуалните стойности на променливата τ_i са условните очаквани стойности на X_i при даден обект на измерване (индивид) u (Стоименова, 2000; Steyer, 2001). Следващото уравнение е конкретизация на уравнение (3):

$$\tau_i = E\left(X_i \middle| U = u\right) \qquad (7)$$

Класическата теория постулира липса на връзка между действителния бал и грешката на измерване, дори и в случите, когато те формират различни наблюдавани балове X_i и X_j. Това допускане се изразява чрез нулевата ковариация (корелация) между тези величини (Lord, 1980):

$$Cov(\tau_i, \varepsilon_j) = 0 \text{ или } Corr(\tau_i, \varepsilon_i) = 0 \quad (8)$$

Средната (очакваната) стойност на грешката на измерване на популационно равнище също е равна на нула:

$$E(\varepsilon_i | U) = 0 \quad (9)$$

1.1.2. Особености на теорията

Класическата тестова теория се гради на поредица от допускания (определяни още и като "аксиоми") относно компонентите в дясната част на основното уравнение (1) и връзката между тях. Основното уравнение, както отбелязват Р. Хамбълтън и Р. Джоунс, съдържа две неизвестни и в този си вид е нерешимо, освен ако не се направят няколко „опростяващи допускания" (Hambleton & Jones, 1993, стр. 255).

Трябва да се отбележи, че голяма част от тези допускания следват от основното уравнение и от дефинициите на двете ненаблюдаеми променливи – действителния бал и грешката на измерването. В този смисъл някои изследователи с основание предпочитат да говорят не за допускания или аксиоми, а за вътрешноприсъщи особености на теорията, които произтичат от разгледаните по-горе уравнения и които са свързани с ненаблюдаваните променливи (Zimmerman, 1976; Steyer, 2001). В допълнение, тези особености могат да бъдат наблюдавани както на индивидуално, така и на популационно равнище (Kline, 2005) и обикновено се отнасят за равнище тестов бал.

1.1.2.1. Допускания относно действителния бал

Разглеждайки СТТ, изследователите разглеждат различен набор от допускания, които могат да бъдат обобщени както следва (виж Hambleton & Jones, 1993; Steyer, 2001; Dawson, 2003; Wiberg, 2004; Kline, 2005; Downing & Haladyna, 2006; Amarnani, 2009 и др.) По-голяма част от тези допускания са валидни както на индивидуално, така и на популационно равнище.

(1) Всеки индивид се характеризира с определен тестов бал, който отразява действителното равнище на измерваната характеристика (личностова черта или състояние, умствена способност, нагласа, интереси).

(2) Измерваната характеристика на индивида е величина, която не може да бъде наблюдавана пряко.

(3) Действителният бал (τ), заедно със случайна грешка на измерването (ε), са адитивни компоненти на наблюдавания бал (X).

(4) Действителният бал на индивида е устойчива, стабилна величина (параметър), която не се променя при многократно измерване с един и същи тест или с раз-

лични (еквивалентни, паралелни) форми на теста (*tau-equivalent/ true-score; equivalent measurement*).

(5) На популационно равнище действителният бал формира интервална скала и е нормално разпределена величина.

(6) При многократни измервания на даден индивид на-добрата оценка на действителния бал е математическото очакване на наблюдаваните тестови балове.

1.1.2.2. Допускания относно грешката на измерването

(1) Грешката на измерването е ненаблюдаема пряко, несистемна и случайна. Това се отнася както за индивидуалните измервания на ниво тестов въпрос, така и ниво тестова бал, където се разглежда като отклонение на наблюдавания от действителния бал (DeVellis, 2003; Downing & Haladyna, 2006).

(2) Грешката на измерването е разпределена нормално, с нулево математическо очакване на индивидуално и на популационно равнище. Това се отнася и до грешката, асоциирана с всеки отделен въпрос. Нейното математическо очакване на популационно равнище също е равно на нула (Hambleton & Jones, 1993; DeVellis, 2003; Kline, 2005).

(3) Грешките на измерване са независими и не корелират помежду си. Това е валидно и на ниво тестов въпрос – грешките при един айтем не корелират с грешките при друг айтем (Hambleton & Jones, 1993; DeVellis, 2003; Kline, 2005). В по-обобщен вид допускането е, че всяко конкретно измерване е съпроводено със специфична грешка и тя е независима от грешката, която може да се допусне при друго измерване на същия конструкт. Това допускане е аналогично на допускането за локална независимост в IRT (Embretson & Reise, 2000, стр. 227).

(4) Тъй като действителната оценка не е известна, не е известен и размерът на грешката. Следователно е възможно само обща оценка на грешката на измерването (Weiner, 2003).

(5) Стандартната грешка на измерване на тестово равнище е еднаква за всички индивиди, т. е. консистентна за всички нива на измерваната характеристика. Независимо от наблюдавания тестов бал, стандартната грешка за всяко негово равнище е еднаква (Lord, 1984; Hambleton, Swaminathan & Rogers, 1991; Michell, 1999; Embretson & Reise, 2000; DeVellis, 2003; Kline, 2005).

1.1.2.3. Допускания относно връзката между действителната стойност и грешката на измерването

(1) Действителният тестов бал (*true score*) и грешката на измерването (*error score*) са независими и не корелират помежду си (*assumption of independence*) (Hambleton & Jones, 1993; DeVellis, 2003; Kline, 2005; Downing & Haladyna, 2006).

(2) Сумата от дисперсиите на действителния бал и на грешката на измерването

е равна на дисперсията на наблюдавания бал (Weiner et al., 2003; Kline, 2005)

1.1.2.4. Допускания относно формата на разпределението на действителния бал, размерността на латентната структура и локалната независимост

Допусканията относно формата на разпределението на действителния бал, размерността на латентната структура и локалната независимост на отговорите не са част от канона на Класическата теория. Независимо от това могат да бъдат привлечени множество доводи от общотеоретичен и прагматичен характер, че тези особености са необходими условия за осъществяване на надеждни измервания в рамките на Класическата тестова теория.

(1) Нормалност на разпределението

Както беше отбелязано, допускането за нормалност на разпределението на тестовия бал обикновено не се разглежда като част от изискванията на Класическата теория. Много изследователи обаче са склонни да приемат това допускане като „съвсем разумно" (Bock & Moore, 1986, стр. 37). Т. Клайн обобщава това мнение като „общо базово допускане при измерването на всяка индивидуална черта", независимо от това дали тя е личностова характеристика, когнитивно, социално или моторно умение, тъй като „тази черта е нормално разпределена в популацията". Затова е важно да се изследва степента, в която това допускане е удовлетворено във всяка извадка от данни (Kline, 2005, стр. 22). Същият автор отбелязва, че на популационно равнище действителният бал формира интервална скала и е „нормално разпределен" (ibid., стр. 94).

За някои от авторите предварителното условие за нормалност е свързано с наблюдаваната (анализираната) променлива (Bock & Moore, 1986; Weiner et al., 2003), с „това, което измерваме" (Boyle & Fisher, 2007, стр. 47). Други отнасят това изискване директно към действителния тестов бал (*true score*) на популационно равнище (Kline, 2005).

Изискването за нормалност на разпределението на дълбинната променлива произтича от обстоятелството, че (почти) всички статистически методи, които се използват в рамките на СТТ, експлоатират математическите свойства на нормалното разпределение. Както основните описателни статистики, така и стандартните грешки на измерването и доверителните интервали се основават на това разпределение (Brennan, 2001).

Класическият индекс на дискриминативна сила на въпросите D се основава на подхода на Т. Кели за определяне на обема на екстремните „силна" и „слаба" групи, които обхващат съответно горните и долните 27% от разпределението на суровия тестов бал. Този подход се основавана анализа на съвместното двумерно нормално разпределение на наблюдавания бал (критериална величина, въз основа на която се определят двете екстремни групи) и действителния тестов бал, отразяващ някаква спо-

собност. Т. Кели доказва, че 27-процентовият обем на екстремните групи осигурява максимизиране на разликата между тях (т. е. между средните стойности на действителния бал на двете групи) по отношение на измерваната способност само ако съвместното двумерно разпределение е нормално (Kelley, 1939, по McCabe, 1978; Kline, 2005).

Бисериалният r_b или точково-бисериалният коефициент на корелация r_{pb}, които са специален случай на коефициента на Пиърсън на смесените произведения, се използват, наред с класическия дискриминативен индекс, като статистическа мярка за оценка на дискриминативната сила на въпросите. Тези коефициенти (и други модификации като φ и ρ на Спиърмън, както и изходният коефициент r на Пиърсън), се базират на допускането, че латентната променлива, която е представена в дискретна дихотомна скала 0/ 1, има непрекъснато разпределение и това разпределение е нормално (Kline, 2005; Калинов, 2010). Ако обаче това разпределение съществено се отклонява от нормалното, по-често използваният бисериален коефициент на корелация r_b може да надхвърли граничната си максимална стойност от ± 1.00 (Калинов, 2010).

В по-общ план корелацията е основна мярка в различни линейни модели и в частност при множествената регресия, при анализа на главни фактори и главни компоненти (включително на дихотомизирани айтеми), при моделирането на структурни уравнения и др. Тя е в основата на различните методи за оценката на надеждността на тестовите резултати, а и на много от методите за оценка на тяхната валидност. Прилагането на параметричните корелационни методи се основава на няколко допускания, едно от които е за нормалност на разпределението на корелираните променливи (Kline, 1998; Embretson & Reise, 2000; Reynolds & Kamphaus, 2003; Kline, 2005).

Всички стандартни скали за трансформация на суровия тестов бал, каквито са стандартните (стандартизирани) z-стойности (нормализирани балове), процентили, *stens*, *stanines*, T-балове и др., почиват на нормалното разпределение (Kline, 1998; Reynolds & Kamphaus, 2003; Kline, 2005; Boyle & Fisher, 2007). Преобразуването на суровия тестов бал представлява линейна трансформация на стойностите на едно нормално разпределение в стойности на друго нормално разпределение.

Нормалното разпределение на тестовия бал е необходима предпоставка и при прилагането на някои други, допълнителни статистически анализи, особено при различните параметрични тестове (например t-test на Стюдънт, F-test на Фишер), базирани на съпоставяне на средни стойности. Относително слаби отклонения от нормалната крива, особено нейната асиметричност, могат да намалят мощността на тези тестове (Weiner et al., 2003). При нормално разпределени величини вероятността, асоциирана със статистическите тестове, е точна оценка на вероятността, свързана с нулевата хипотеза. Но когато те не са нормално разпределени, тогава оценената вероятност може да бъде прекалено консервативна или либерална (Weiner et al., 2003).

(2) Едномерност на латентната структура

Традиционно класическите тестови модели не поставят ограничението за едномерност на тестовите данни. Достатъчно е латентната структура, с каквато и размерност да е тя, да бъде еднаква при всички паралелни форми на съответния тест (Hambleton & Jones, 1993).

Тук обаче е уместно да припомним двата основни постулата на Л. Л. Търстоун за психологическите измервания, които постулати, както отбелязва авторът, за разлика от случаите на физически измервания, трябва да бъдат формулирани експлицитно. Първият от тях гласи, че „без значение какъв може да бъде обектът на измерване, *измерването описва само един атрибут на обекта*". Ако е необходимо по-пълно описание на този обект, могат да се измерят няколко от неговите атрибути. Друг постулат, „който лежи в основата на всички измервания е, че *измервания атрибут е винаги едномерен*". И тъй като не всички атрибути са едномерни, не всички могат да бъдат измерени (Thurstone, 1929, стр. 158-159, подчертаването е на автора). Много автори смятат, че определянето на размерността на скритото пространство е „критичен въпрос при моделирането на тестовите данни", независимо от това дали се обсъжда прилагането на едномерен или многомерен модел (Embretson & Reise, 2000, стр. 227; de Ayala, 2009).

Най-съществено ограничение пред използването на вътрешната консистентност като мярка за надеждността на теста е изискването за едномерност на тестовия бал. Коефициентът $alpha$ на Кронбах е особено чувствителен към всяко отклонение от едномерността на тестовите въпроси в случите, в които те са ориентирани към повече от един конструкт (Cortina, 1993; Abedi, 1996).

(3) Локална независимост

Някои автори приемат локалната независимост на въпросите за „фундаментално допускане във всички модели в рамките на теорията на измерването" (Downing & Haladyna, 2006, стр. 297). Отсъствието на локална независимост може да доведе до надценяване на надеждността на тестовия бал. Такава ситуация може да бъде наблюдавана при тестове, съдържащи контекстово-зависими групи от въпроси - корелациите между въпросите в дадена група обикновено са много по-високи, отколкото с въпросите вън от групата. Високите корелации между контекстово-зависими въпроси от своя страна водят до „изкуствено" повишаване на общото равнище на надеждността на теста (Wainer & Thissen, 1996)

1.1.3. Модели на CTT

Обикновено CTT се разглежда като единна теория, но когато някои от теоретичните параметри като надеждност, дисперсия на действителния бал или на грешката на измерване се оценяват чрез емпирични статистики като средни стойности, дисперсии

или корелации на наблюдаваните балове, е необходимо да се формулират допълнителни допускания, които формират различни модели в рамките на CTT (Steyer, 2001; DeVellis, 2003). Различните модели на тестове в рамките на Класическата теория се определят като такива чрез отношенията им с други тестове, които изпълняват същите функции.

Най-често се прилагат следните три едномерни модела, чрез които се изразяват отношенията между които и да е два теста X_i и X_j от съвкупността $X_1, X_2, \ldots, X_m$: (1) паралелни тестове (*parallel tests)*, (2) основно τ-еквивалентни тестове (e*ssentially* τ-*equivalent tests)* и (3) τ-конгенерични тестове (τ-*congeneric tests)* (Steyer, 2001; DeVellis, 2003).

Допусканията, на които се основават моделите, могат да бъдат обособени в две групи:

(1) Допускания по отношение на действителните балове τ_i и τ_j:

τ-еквивалентност: $\tau_i = \tau_j$ (10)

основна τ-еквивалентност: $\tau_i = \tau_j + \lambda_{ij},\ \lambda_{ij} \in \Re$ (11)

τ-конгенеричност: $\tau_i = \lambda_{ij0} + \lambda_{ij1}\tau_j,\ \lambda_{ij0}, \lambda_{ij1} \in \Re,\ \lambda_{ij1} > 0$ (12)

(2) Допускания по отношение на грешките на измерване ε:

липса на корелация: $Cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$ (13)

равни дисперсии: $Var(\varepsilon_i) = Var(\varepsilon_j)$ (14)

Допусканията от първата група дефинират по различен начин обстоятелството, че два теста X_i и X_j могат да бъдат използвани за оценка на една и съща индивидуална характеристика. Най-силното е допускане (10), в което се предполага пълно равенство между действителните балове на двата теста. Следващото е по-слабо – в него се допуска разлика между двата действителни бала с определена константа λ_{ij}, т.е. линейна корелация с коефициент $b = 1$. Третото допускане, което е най-слабо, също предполага разлика между действителните балове на двата теста, които корелират помежду си линейно с коефициент $b \neq 1$.

Допусканията във втората група се отнасят до отношенията между грешките на измерване в двата теста. Допускане (13) е за липса на корелация между съответните променливи, а (14) е за равенство на техните дисперсии, т.е. двата теста измерват еднакво добре и се характеризират с еднаква прецизност.

Моделите на CTT се формират от различни комбинации между горните допускания (Steyer, 2001).

Паралелните тестове отговарят на условие (10) за τ-еквивалентност, на (13)

за липса на корелация между грешките и (14) за равенството на техните дисперсии. Допускането за τ-еквивалентност предполага наличието на ясно дефинирана индивидуална характеристика (или латентна променлива), която е една и съща за всяка τ-променлива, т. е. „замества" действителните балове на всеки един от паралелните тестове τ_1, τ_2, …, τ_m. Ако означим латентната променлива с η, допускането за τ-еквивалентност ще приеме следния вид:

$$X_i = \eta + \varepsilon_i, \text{ където } \varepsilon_i = X_i - E\left(X_i \middle| U\right) \quad (15)$$

Основно τ-еквивалентните тестове отговарят на условия (11) за основна τ-еквивалентност и (13) за липса на корелация между грешките. Това допускане, както и в предходния модел, също предполага наличието на латентна променлива η, която е отразена в действителния бал на всеки тест:

$$\tau_i = \eta + \lambda_i, \quad \lambda_i \in \Re \quad (16)$$

така че:

$$X_i = \eta + \lambda_i + \varepsilon_i, \text{ където } \varepsilon_i = X_i - E\left(X_i \middle| U\right) \quad (17)$$

τ-*конгенеричните тестове* се определят от допускане (12) за τ-конгенеричност и (13) за липса на корелация между грешките. Това допускане също предполага наличието на латентна променлива η, а действителният бал на всеки тест е нейна линейна функция:

$$\tau_i = \lambda_{i0} + \lambda_{i1}\eta, \quad \lambda_{i0}, \ \lambda_{i1} \in \Re, \quad \lambda_{i1} > 0 \quad (18)$$

или еквивалентното уравнение:

$$X_i = \lambda_{i0} + \lambda_{i1}\eta + \varepsilon_i, \text{ където } \varepsilon_i = X_i - E\left(X_i \middle| U\right) \quad (19)$$

1.1.4. CTT като теория за надеждността

В сърцевината на Класическата тестова теория стои проблемът за надеждността. Сред изследователите има „почти всеобщо съгласие", че надеждността е съществена характеристика на образователните измервания (Colton et al., 1997, стр. 3). Ролята на този конструкт в класическите теоретични модели е толкова важна, че нерядко CCT се разглежда именно като теория за надеждността (*reliability theory*) (DeVellis, 2003).

Нека най-напред да се занимаем с въпроса на кое от явленията, свързани с инструментите за психологическо оценяване, следва да се припише характеристиката "надеждност", или с други думи, кое е надеждно? Често пъти можем да чуем фрази ка-

то „този тест (въпросник, инструмент, метод) е надежден (или недостатъчно надежден)”. В Класическата теория надеждността е характеристика, която се приписва единствено на наблюдавания тестов бал. Тя отразява основния „конфликт” в теорията – в резултат на едно измерване изследователят получава директна информация за наблюдавания тестов бал, от който не се интересува, но не и за действителния бал, от който се интересува. Надеждността винаги се свързва с тестовия бал, получен чрез някакъв измервателен инструмент, при конкретна негова употреба, за конкретна група от изпитани, а не със самия инструмент или процедура (Thompson, 2003).

В специализираната литература могат да бъдат срещнати десетки различаващи се дефиниции на надеждността, в които тя е определяна като "консистентност", "точност", "възпроизводимост", "повторяемост", "сигурност" и "стабилност" на тестовите резултати. Но това са по-скоро "технологични" дефиниции, които описват начините, по които може да се направи оценка на надеждността. По-адекватният подход към надеждността е тя да бъде разглеждана през призмата на действителния бал, в който е фокусиран изследователския интерес.

Надеждността може да се определи като степента, в която наблюдаваният резултат от едно измерване отразява или по-скоро съдържа в себе си действителната стойност на измервания признак. С други думи, това е мярка за степента на сигурност, че наблюдаваният резултат може да се използва за оценка на ненаблюдавания, но единствено важен компонент в основното уравнение на СТТ. Ако се върнем към уравнение (1), ще видим, че действителният бал и грешката на измерването се допълват взаимно в наблюдавания бал. Ако размерът на грешката (т. е. нейният дял в наблюдавания бал) намалява, расте размерът (делът) на действителния бал. В идеалния случай, когато не е допусната грешка и компонентът ε има нулева стойност (т. е. бъде „изваден” от уравнение 1), резултатът от измерването съдържа в себе си само действителния бал τ. И обратно, когато размерът на грешката (т.е. нейният дял в наблюдавания бал) расте, намалява размерът (делът) на действителния бал. В най-лошия хипотетичен случай, ако компонентът τ има нулева стойност (т. е. бъде „изваден” от уравнение 1), резултатът от измерването може да съдържа в себе си само грешка. В този смисъл надеждността може да се разглежда и като степента, в която резултатът от едно измерване е свободен от грешки – надеждността расте с намаляване на размера на грешката.

Макар че представляват основен изследователски интерес, стойностите на τ и ε не могат да бъдат наблюдавани чрез емпирично изследване (Lord, 1980). Възможно е обаче да се направи оценка на техните дисперсии въз основа на случайна представителна извадка от изпитани лица. Т. Доусън предлага една интересна гледна точка към надеждността на измерването. Авторът подхожда към този проблем като при статистическите анализи на зависими променливи – чрез разлагане на наблюдаваната дисперсия на съставните й компоненти. Т. Доусън показва, че моделът на разлагане на

дисперсията на два компонента – обяснена и необяснена дисперсия (дисперсия на грешката), който се прилага например при регресионния или дисперсионния анализ, може да бъде приложен и към наблюдавания бал (Dawson, 2003).

Действително, едно от важните следствия от независимостта между τ и ε е това, че ако на тестиране са подложени група от индивиди, дисперсията на техните наблюдавани балове може да бъде представена като сума от дисперсията на действителните им балове и дисперсията на грешката. Т. Доусън определя компонентите в дясната част на равенство (20) като "надеждна" (*reliable*) и "ненадеждна" (*unreliable*) дисперсия (Lord, 1980; Dawson, 2003; Weiner et al., 2003).

$$Var(X_i) = Var(\tau_i) + Var(\varepsilon_i) \text{ или } \sigma_X^2 = \sigma_\tau^2 + \sigma_\varepsilon^2 \quad (20)$$

Посредством компонентите в горните уравнения може да се представи и корелацията между наблюдавания и действителния бал, която се прилага като основен статистически подход за изследване на надеждността. Силата на линейна връзка между тези компоненти на основното уравнени се представя чрез следната формула:

$$\rho(X,\ \tau) = \frac{\sigma_\tau^2}{\sqrt{(\sigma_\tau^2 + \sigma_\varepsilon^2)\ \sigma_\tau^2}} \quad (21)$$

Чрез използването на коефициента на линейна корелация в Класическата теория се постулира имплицитно наличието на линейна връзка между наблюдавания и ненаблюдавания тестов бал. Но надеждността на наблюдавания бал не е просто корелацията му с ненаблюдавания бал. Като оценка на този показател се използва квадратът на корелацията между тези две променливи (Lord, 1980; Стоименова, 2000; DeVellis, 2003; Downing & Haladyna, 2006), често обозначаван в статистиката като коефициент на детерминация (r^2):

$$\rho^2(X,\ \tau) = \left[\frac{\sigma_\tau^2}{\sqrt{\left(\sigma_\tau^2 + \sigma_\varepsilon^2\right)\sigma_\tau^2}}\right]^2 = \frac{\sigma_\tau^2}{\sigma_X^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_X^2} \quad (22)$$

Следователно в най-абстрактна форма надеждността на измерването се представя като отношение между дисперсията на действителния тестов бал и дисперсията на общия наблюдаван тестов бал (Weiner et al., 2003). Използването на коефициента на детерминация има особен статистически смисъл – той показва каква част от дисперсията на наблюдавания бал се дължи на (и може да бъде обяснена чрез) дисперсията на действителния бал. Когато ρ^2 се приближава към 1.00, неизвестната стойност τ се приближава към наблюдаваната стойност X, която може успешно да се използва

вместо нея. Оттук възможно най-кратката дефиниция на надеждността може да се формулира така: това е частта от дисперсията на наблюдавания бал, която се дължи на действителния бал.

Разработени са няколко подхода за оценка на ефекта на грешката върху тестовите резултати, т.е. на тяхната надеждност, които се основават на метода, предложен от Л. Л. Търстоун, според когото надеждността на един тест се заключава в корелацията със самия него (Thurstone, 1931a). Образното твърдение на Л. Л. Търстоун следва да се разбира в смисъл, че надеждността (корелацията между наблюдавания и действителния бал) може да бъде изразена чрез корелацията между наблюдаваните балове на два паралелни теста X_i и X_j. Математическият израз на този метод за оценка на надеждността се задава чрез формулата[3]:

$$\mathrm{Re}\,l(X,\ \tau)=\rho(X_i,\ X_j) \qquad (23)$$

където:

$Rel(X, \tau)$ - надеждност на наблюдавания бал X

$\rho(X_i, X_j)$ - корелация между наблюдаваните балове X_i и X_j, получени при повторно използване на същия тест или на паралелен тест

В СТТ се разглеждат три типа (три източника) на грешки, съответно три подхода за измерване на надеждността (Crocker & Algina, 1986), които в литературата понякога некоректно се определят като видове надеждност:

(1) грешка на измерване, която се дължи на момента на използване на теста (*stability*);

(2) грешка на измерването, която се дължи на тестовата форма (*equivalence*);

(3) грешка на измерването, която се дължи на айтемите (*internal consistency*).

1.1.5. Оценяване на надеждността

1.1.5.1. Оценка на надеждността чрез повторно използване на един и същи тест (test-retest reliability)

Методът се състои в двукратно използване на една и съща тестова форма, с една и съща извадка от лица, но в различни моменти от време. Пиърсъновият продукт-момент коефициент на корелация между двете серии от наблюдавани тестови балове, получени при първото и второто измерване, е мярка за надеждността на резултатите, която при този подход се изразява като тяхна стабилност, устойчивост във времето.

[3] Учудващо е, че при безспорната важност на понятието за надеждност, в литературата няма общоприет начин за неговото обозначаване. Използват се различни символи, включително и "*r*" или "ρ", с които обикновено се обозначава коефициентът на корелация. Поради това тук и по-нататък, където е необходимо, ще използваме обозначението "*Rel*".

Следователно тест-ретестовата надеждност е мярка за степента, в която резултатите от едно тестиране могат да бъдат генерализирани за различни случаи на употреба на този тест. Някои изследователи препоръчват като коефициент на стабилност (*coefficient of stability*) да се използва квадратът на коефициента на корелация (Dawson, 2003).

Методът се основава на допускането относно действителния бал, съгласно което той е устойчива величина, която не се променя в интервала между двете тестирания. Следователно лицата биха отговаряли консистентно при първото и второто тестиране, т. е. наблюдаваните тестови балове на всяко лице при първото и второто тестиране биха съдържали един и същ действителен бал, характеризиращ това лице. Колкото по-малки са грешките на измерването, толкова по-висока би била "чистотата" на наблюдаваните балове, а оттам и корелацията между тях. Следователно, измервайки стабилността на наблюдаваните резултати през определен период от време, се измерва тяхната надеждност.

От съществено значение за големината на коефициента на стабилност е дължината на времевия интервал, през който се извършват последователните тестирания. Ако той е по-кратък, коефициентът на стабилност е по-висок и обратно – при по-продължителни периоди между двете тестирания неговата стойност намалява (Pedhazur & Schmelkin, 1991; Анастази и Урбина, 2001). Поради това, ако измерим надеждността на един тест многократно, през различни времеви интервали, бихме могли да получим множество от тест-ретестови коефициенти на надеждност (Анастази и Урбина, 2001), което прави тази мярка за надеждност твърде "ненадеждна".

Ефектът на времевия интервал върху оценката на надеждността може да бъде обяснен с действието на два фактора. При по-кратък времеви интервал вероятността от промяна в действителните балове на изпитваните лица е по-малка, а вероятността от припомняне на отговорите, които са дали при първото тестиране – по-голяма. При тази ситуация двата фактора действат в една посока – към по-висока възпроизводимост (стабилност) на резултатите. И обратно – при по-голям времеви интервал вероятността от промяна в действителните балове на изпитваните (поради естествените процеси на забравяне или научаване) е по-висока, а вероятността от припомняне на отговорите, които изпитваните са дали при първото тестиране – по-малка. И при тази ситуация двата фактора действат в една посока – към по-ниска възпроизводимост (стабилност) на резултатите. В идеалния случай двете измервания би следвало да се направят през кратък период от време (непосредствено едно след друго) при условие, че при второто тестиране изпитваните не помнят отговорите, които са дали при първото.

1.1.5.2. Оценка на надеждността чрез използване на алтернативни тестове (alternate/ equivalent forms reliability)

За да се избегнат негативните ефектите от дължината на времевия интервал, при този метод се използват две алтернативни тестови форми, с една и съща извадка от лица, в приблизително едно и също време. Алтернативни са два (или повече) варианта на един тест, които се разглеждат като взаимно заменими, т.е. предназначени са за постигане на едни и същи изпитни цели, измерват един и същи теоретичен конструкт, разработени са по един и същи тестов план и се провеждат по един и същи начин. За алтернативните форми може да се мисли, че са съставени от един и същи пул от въпроси, разделени на две половини по случаен начин.

Пиърсъновия продукт-момент коефициент на корелация между резултатите от двете тестирания, обозначаван като индекс на надеждност (*reliability index*), е мярка за еквивалентността на двете форми (DeVellis, (2003; Dawson, 2003). Според някои изследователи, този индекс трябва да се повдигне на квадрат, за да стане пълноценна мярка за надеждността на измерването, т.е. коефициент на надеждност (Gronlund & Linn, 1990).

Идеални „кандидати" за алтернативни тестове са паралелните тестове, които отговарят на условие (10) за τ-еквивалентност, на (13) за липса на корелация между грешките и (14) за равенството на техните дисперсии. Ако моделът е по-слабо консервативен, например тестовете са основно τ-еквивалентни, надеждността се определя чрез ковариацията между наблюдаваните балове (Steyer, 2001):

$$\mathrm{Re}\,l(X_i) = \frac{Cov(X_i,\, X_j)}{Var(X_i)}, \quad i \neq j \tag{24}$$

При основно τ-еквивалентни тестове $X_1, X_2, \ldots, X_m$ за оценка на надеждността на сумарната скала $S = X_1 + \ldots + X_m$ може да се използва и коефициентът α на Кронбах, който ще бъде разгледан детайлно по-нататък в текста:

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum_{i=1}^{m} Var(X_i)}{Var(S)}\right) \tag{25}$$

Оценката на надеждността чрез използване на алтернативни тестови форми също има свои слабости, които се коренят в практическата трудност за съставяне на такива форми. Тъй като тестирането се извършва (почти) едновременно, с едни и същи лица, но с форми, съставени от различни въпроси, източникът на ненадеждност трябва да се търси в разликите между извадките от въпроси.

1.1.5.3. Оценка на надеждността чрез изследване на вътрешната консис-

тентност на въпросите (internal consistency reliability)

Поради практическите трудности, произтичащи от прилагането на първите два метода за оценка на надеждността, повечето изследователи не биха прибягнали към прилагането на един и същи тест двукратно или към съставянето на две алтернативни форми на теста. Изследването на вътрешната консистентност се основава на еднократното прилагане на дадена тестова форма върху дадена извадка от лица. Въпреки това в основата му лежи същата идея – изследване на корелацията между две тестови форми, но при този метод те се конструират чрез виртуалното разделяне на изходния тест на (равни) части, които се разглеждат като алтернативни една на друга.

Първият подход за оценка на надеждността чрез вътрешната консистентност на въпросите е чрез „физическото" разделяне на теста на две половини (*split-half reliability*). По определена схема (чрез случаен подбор, разделяне на четни и нечетни и др.) въпросите в теста се разделят на две равни части, на два производни "теста", които трябва да отговарят на условието за паралелност. След това се определя тестовият бал на всяко лице по всяка половина на теста и се изчислява корелацията между двете серии наблюдавани балове, която се разглежда като мярка за съгласуваността (*agreeability*) между двете части на теста. Повдигнат на квадрат, корелационният коефициент служи за оценка на надеждността на едната половина от теста, но не и за теста като цяло (Dawson, 2003).

За оценка на надеждността на целия тест, въз основа на получената корелация между неговите половини, се прилагат подходът, разработен независимо от Ч. Спиърмън и У. Браун в далечната 1910 г. Тяхната "предсказваща формула" (*Spearman-Brown prediction/ prophecy formula, Spearman-Brown split-half reliability coefficient*) обвързва надеждността на теста с неговата дължина, т.е. с броя на въпросите (Thorndike, Cunningham, Thorndike & Hagen, 1991). Формулата се използва за решаването на два типа задачи: (1) за определяне ("предсказване") на надеждността на бъдещ тест чрез добавяне (или изваждане) на фиксиран брой въпроси с характеристики, аналогични на тези от съществуващия тест и (2) за определяне ("предсказване") на необходимия брой въпроси, които трябва да бъдат добавени към съществуващия тест за достигане на определена (желана) надеждност на бъдещия тест. Поради това методът понякога се нарича "усилваща" ("*step up*") формула, чийто общ вид е следният (по Weiner et al., 2003):

$$\mathrm{Re}\, l_{S-B} = \frac{n.\rho_{xy}}{1+(n-1).\rho_{xy}} \qquad (26)$$

където:

$Rel_{S\text{-}B}$ - очаквана стойност на коефициента на надеждност на Спиърмън-Браун

n - отношение между броя на въпросите в бъдещия и в съществуващия тест

ρ_{xy} - коефициент на надеждност на съществуващия тест

Формулата на Спиърмън-Браун може да се използва за определяне на надеждността на тест, разделен на две половини, тъй като броят на въпросите от едната половина, надеждността на която е известна, практически се удвоява с въпросите от другата половина. В този случай се използва следната опростена предсказваща формула:

$$\mathrm{Re}\, l_{S-B} = \frac{2.\rho_{xy}}{1+\rho_{xy}} \qquad (27)$$

Подходът на разделяне на теста на половини също има свой „вроден" недостатък, който е повече от очевиден – по правилата на комбинаториката могат да бъдат формирани множество двойки от виртуални половини, съдържащи различни комбинации от въпроси. Поради това е напълно възможно получените двойки половини на теста да се различават една от друга, отдалечавайки се от изискването за паралелност. Оттук и оценките на вътрешната консистентност, получени след прилагането на един или друг начин за разделяне на въпросите в теста, могат да варират значително (Sax, 1989).

Съществен напредък в посока към преодоляване на проблема с множеството възможни оценки на надеждността са процедурите за оценка на вътрешната консистентност, предложени от Дж. Кюдър и М. Ричардсън, които разработват серия от индекси (Kuder & Richardson, 1937, по Cronbach, 1951 и Weiner et al., 2003). Математическия израз на най-използвания сред тях, известен като "Формула 20 на Кюдър-Ричардсън" (*Kuder-Richardson Formula 20, KR-20*), е следният:

$$\mathrm{Re}\, l_{KR-20} = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{\sigma_X^2}\right) \qquad (28)$$

където:

k - брой на въпросите в теста

p_j - дял на изпитаните, отговорили правилно на j-тия въпрос

q_j - дял на изпитаните, отговорили неправилно на j-тия въпрос

σ_X^2 - дисперсия на общия (композитен) тестов бал

С. Лайърли разглежда формулата на Кюдър-Ричардсън като генерализиран Спиърмън-Браунов коефициент. Той доказва, че основната й предпоставка - матрицата на корелациите между айтемите има ранг 1, т.е. всички айтеми измерват една и съща латентна променлива - е достатъчна, за да обоснове многобройните допускания, използвани от различни автори, разработили нейни деривати (Lyerly, 1958; Vehkalahti,

2000). Л. Кронбах от своя страна установява, че KR-20 е начин за определяне на средната стойност на коефициентите на Спиърмън-Браун, изчислени за всички възможни половини на теста (Cronbach, 1951).

Като недостатък на формулата на Кюдър-Ричардсън се посочва обстоятелството, че тя е приложима единствено за дихотомични скали. Действително, ако се върнем на формула (28), ще видим, че авторите са базирали своя подход върху разпределението на правилните и неправилните отговори (скорирани с 1/ 0) на изпитаните лица на всеки тестов въпрос. Л. Дж. Кронбах, специалист по психология на образованието, прави още една стъпка към генерализиране на методите за оценка на надеждността, разработвайки един от най-популярните индикатори за вътрешната консистентност на въпросите в измервателния инструмент (Cronbach, 1951). Коефициентът *alpha* (Cronbach's α) се разглежда като наследник на няколко предходни мерки за вътрешна консистентност – не само на формулата на Кюдър-Ричардсън 20, но и на серията от мерки за определяне на долната граница на надеждността (*lambda* -1 до *lambda* - 6), разработени от Л. Гутман (Guttman, 1945). Откакто коефициентът α на Кронбах е станал неразделна част от психометричния инструментариум, останалите мерки за вътрешна консистентност се използват по-ограничено, тъй като няма данни за някакви техни предимства пред *alpha*. KR-20 дори се разглежда като негова производна, подходяща за дихотомични въпроси, докато коефициентът α може да се прилага както за дихотомични, така и за континуални и Ликертови скали (Crocker & Algina, 1984).

Коефициентът α на Кронбах е пряко развитие на методите за оценка на надеждността чрез разделяне на теста на две половини, с които споделя един и същи концептуален подход. Той може да бъде разгледан по различен начин: като оценка на математическото очакване на квадрата на корелацията между двойки тестове, извлечени по случаен начин от пул от въпроси, сходни с тези в съществуващия тест; като средна стойност (математическото очакване) на коефициентите на надеждност, получени от комбинацията между всички възможни половини на съществуващия тест; като мярка за съгласуваността между въпросите и общия тестов бал и т.н. Психометричното значение на коефициента α е да служи като оценка на онази част от дисперсията на наблюдавания тестов бал, която може да бъде обяснена чрез (или която той споделя с) ненаблюдавания действителен бал (DeVellis, 2003).

Математически коефициентът α може да бъде представен чрез различни еквивалентни формули, най-често използваната сред които е следната (DeVellis, 2003):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma_j^2}{\sigma_X^2}\right) \qquad (29)$$

където:

k - брой на компонентите (въпроси, субтестове) в теста

σ_j^2 - дисперсия на *j*-тия въпрос

σ_X^2 - дисперсия на общия (композитен) тестов бал

Друг математически израз на тази мярка за вътрешна консистентност е формулата за стандартизирания коефициент α (*standardized item alpha*), който се изчислява след предварително стандартизиране на въпросите:

$$\alpha = \frac{k.\bar{r}}{[1+(k-1).\bar{r}]} \qquad (30)$$

където:

k - брой на компонентите (въпроси, субтестове) в теста

$\bar{r}$ - средната стойност на коефициентите на корелация на Пиърсън между компонентите

или:

$$\alpha = \frac{k.\bar{c}}{[\bar{v}+(k-1).\bar{c}]} \qquad (31)$$

където:

k - брой на компонентите (въпроси, субтестове) в теста

$\bar{c}$ - средната ковариация между компонентите (*average inter-item covariance*)

$\bar{v}$ - средната вариация (*average variance*)

1.1.6. Тестови статистики

Голяма част от психометричните анализи, основани на Класическата теория, са фокусирани по-скоро върху оценяване на знанията и уменията на равнище тест, отколкото на равнище тестов въпрос. Независимо от това, наред с разнообразието от скалови статистики, предназначени за оценка на цялостния тест, Класическата теория включва и редица статистики, предназначени за оценка на характеристиките на отделните въпроси, а и на още по-ниско ниво – на алтернативните отговори. При все това по отношение на тестовите въпроси теорията е сравнително бедна, тъй като не съдържа модел, който да обвързва способността на изпитваните лица с успеха им при решаване на отделния въпрос.

- *Брой на въпросите в теста или субтеста* (k).
- *Брой на изпитаните лица* (N).
- *Среден бал* $\left(\overline{X}\right)$. Изчислява се като среден брой на въпросите в теста, на които изпитаните лица са отговорили правилно, т.е. като средноаритметична стойност на

наблюдавания бал. Може да се използва като абсолютен индекс за трудността на целия тест.

• *Медиана* на разпределението на наблюдавания бал. Това е 50-ия процентил, който разделя тестовите балове на две равни части. Половината от изпитаните лица имат тестови балове, равни на (или по-ниски от) медианата.

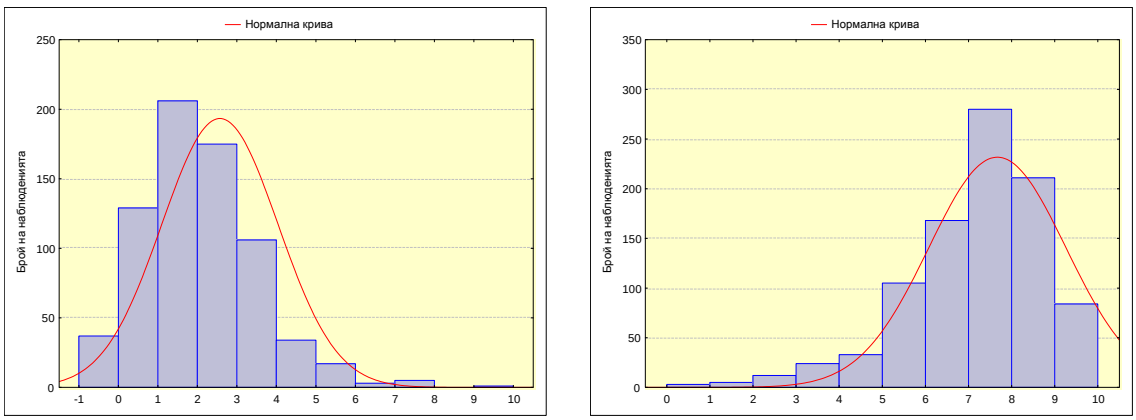
• *Дисперсия* $\left(s_x^2\right)$ на разпределението на наблюдавания бал, която служи за оценка на разсейването на индивидуалните балове около средната им стойност и дава информация за хомогенността на лицата в извадката по отношение на техните способности. Изчислява се по стандартната статистическа формула.

• *Стандартно отклонение* (s_x) на разпределението на наблюдавания бал, което също служи за оценка на разсейването на индивидуалните балове около средната им стойност и носи аналогична информация. Изчислява се по стандартната статистическа формула, като втори корен от дисперсията.

• *Асиметрия* на разпределението на наблюдавания бал, която носи информация за неговата форма. При положителна стойност на този параметър разпределението е асиметрично, изтеглено в лявата част на скалата, в посока към по-ниски тестови балове, а при отрицателна стойност - изтеглено към дясната й част. При нулева стойност тестовите балове са разпределени симетрично около центъра (средната стойност) на разпределението.

Асиметрията на разпределението има отношение към т.н. "подов" и "таванен" ефект. Първият се наблюдава при положително асиметрични разпределения, в които голяма част от изпитаните лица имат ниски балове.

Фигура 1. Разпределения на тестовия бал с положителна и отрицателна асиметрия



В този случай тестът дискриминира слабо лицата, намиращи се в лявата част на скалата. И обратно, при отрицателно асиметрични разпределения голяма част от изпитаните лица имат високи тестови балове и в този случай тестът дискриминира слабо лицата, намиращи се в дясната част на скалата.

• *Ексцес* на разпределението на наблюдавания бал. Това е статистическа мярка за изпъкналостта на централната част на разпределението. При положителна стойност емпиричното разпределение е повече изпъкнало в сравнение с нормалното разпределение, а при отрицателна стойност то е по-плоско. При нулева стойност ексцесът на емпиричното разпределение съвпада с този на нормалното.

• *Минимум* (x_{min}), или най-ниският тестов бал, получен от едно (или повече) от изпитаните лица.

• *Максимум* (x_{max}), или най-високия тестов бал, получен от едно (или повече) от изпитаните лица.

• *Надеждност на резултатите* от измерването. Показател за стабилността, за еквивалентността или за консистентността на тестовия бал. Оценява се по някой от начините, описани по-горе.

• *Стандартна грешка на измерването* (SEM). Това е оценка на стандартното отклонение на грешките на измерването в наблюдаваните балове. Изчислява се по следната формула:

$$SEM = \sigma\sqrt{1 - \mathrm{Re}\,l} \qquad (32)$$

където:

σ - стандартно отклонение на наблюдавания тестов бал

Rel - надеждност на теста

• *Средна трудност/ дял на правилните отговори* ($mean\ P$). Това е мярка за отношението между правилните и неправилните отговори на всички въпроси в теста. Изчислява се като отношение между броя на правилните отговори и всички възможни отговори.

$$mean\ P = \frac{k_r}{k\,.\,N} \qquad (33)$$

където:

k_r - брой на правилните отговори

k - брой на въпросите

N - брой на изпитаните лица

• *Средна корелация между въпросите и тестовия бал* (*Mean item-total correlation*). Изчислява се като средна стойност на точково-бисериалните коефициенти на корелация на всички въпроси в теста с тестовия бал.

• *Среден бисериален коефициент на корелация* (*Mean biserial*). Изчислява се като средна стойност на бисериалните коефициенти на корелация на всички въпроси в теста с тестовия бал.

• *Брой на лицата в "слабата група"* (N_{low}). Тази група се формира (обикновено) от 27% от общия брой на изпитаните лица, които имат най-ниски тестови балове, ако всички изпитани лица са подредени в низходящ ред.

• *Максимален бал* (*слаба група*) ($x_{low,\ max}$). Това е максималният тестов бал, получен от едно или повече изпитани лица, принадлежащи към "слабата група". Определя се чрез интервала, който съдържа 27-мия процентил.

• *Брой на лицата в "силната група"* (N_{high}). Формира се (обикновено) от 27% от общия брой на изпитаните лица, които имат най-високи тестови балове, ако всички изпитани лица са подредени в низходящ ред.

• *Минимален бал (силна група)* ($x_{high,\ min}$). Това е минималният тестов бал, получен от едно или повече изпитани лица, принадлежащи към "силната група". Определя се чрез интервала, който съдържа 73-мия процентил.

• *Корелации между субтестовите скали*. Когато тестът съдържа две или повече субскали, се изчислява Пиърсъновия продукт-момент коефициент на корелация r между всеки две от тях.

1.1.7. Статистики на въпросите

В рамките на разглежданата теория всеки въпрос се характеризира с три количествени индекса: трудност, дискриминативна сила и индекс на налучкване на верния отговор.

• *Трудността на въпроса* (*difficulty, proportion correct*) е мярка за това как изпитаните са се справили със задачата, поставена в него. Във формален план въпросът "как" се трансформира в "колко" от изпитаните са отговорили правилно на даден въпрос и колко са дали неправилен отговор. Колкото по-малко изпитани лица са отговорили правилно, толкова по-труден е въпросът. Математическият израз на тази концепция е следният:

$$p_j = \frac{n_r}{N} \tag{34}$$

където:

p_j - трудност на j-тия въпрос

n_r - брой на изпитаните лица, отговорили правилно на въпроса

N - общ брой на изпитаните лица

Тъй като индексът на трудността изразява отношение (пропорция) между две количества, той може да приема стойности в интервала $0.00 \leq p \leq 1.00$. Трудността е намаляващ индекс – колкото по-ниска е неговата стойност, толкова по-труден е въпросът. В най-лошия случай, когато нито един от изпитаните не е дал правилен отговор на даден въпрос, стойността на индекса $p_j = 0.00$. И обратно, ако всички изпитани са отго-

ворили правилно, това е въпрос с минимална трудност (или максимална „лесната”) с индекс p_j = 1.00.

• *Дискриминативната сила* (*discrimination index*) е следващата, не по-малко важна характеристика на въпросите. За разлика от трудността, която е "очевидно" понятие, дискриминативната сила е „невидима” характеристика, която изразява „способността” на въпросите да разделят индивидите с високи и ниски равнища на способности, като “пропускат” силните и “дискриминират” слабите. Един въпрос с висока дискриминативна сила „позволява” да бъде решен правилно от по-подготвените и „не позволява” същото на по-малко подготвените.

Класическият дискриминативен индекс се основава на разликата между дела на изпитаните лица в силната и слабата група, отговорили правилно на съответния въпрос, и се изчислява по следната формула:

$$D_j = p_{r,\,high} - p_{r,\,low} \tag{35}$$

където:

D_j – дискриминативна сила на j-тия въпрос

$p_{r,\,high}$ - делът на изпитаните лица от силната група, отговорили правилно на j-тия въпрос

$p_{r,\,low}$ - делът на изпитаните лица от слабата група, отговорили правилно на j-тия въпрос

Дискриминативният индекс може да приема стойности в интервала $-1.00 \leq D \leq 1.00$, като отрицателните стойности са признак за лошо, неприемливо качество на съответния въпрос.

Като алтернативна и по-ефективна оценка на дискриминативната сила на въпроса се използва и точково-бисериалният (или бисериалният) коефициент на корелация между отговорите на изпитаните лица на съответния въпрос и бала им по скалата, към която той принадлежи. Точково-бисериалният коефициент е мярка за корелацията между две променливи, едната от които е измерена в дихотомична скала (отговорите на изпитаните лица на съответния въпрос са скорирани 1/0), а другата – в метрична скала (скаловият им бал). Коефициентът е вариант на Пиърсъновия продукт-момент коефициент на корелация и се изчислява по следната опростена формула (по Глас и Стэнли, 1976):

$$r_{pb} = \frac{\overline{X}_r - \overline{X}_w}{s_x}\sqrt{\frac{n_r n_w}{N(N-1)}} \tag{36}$$

където:

r_{pb} - точково-бисериален коефициент на корелация

$\overline{X}_r$ - средната стойност на тестовия бал на изпитаните лица, отговорили правилно на въпроса

$\overline{X}_w$ - средната стойност на тестовия бал на изпитаните лица, отговорили неправилно на въпроса

s_x - стандартно отклонение на всички тестови балове

n_r - брой на изпитаните лица, отговорили правилно на въпроса

n_w - брой на изпитаните лица, отговорили неправилно на въпроса

$N = n_r + n_w$ - общ брой на изпитаните лица

Точково-бисериалният коефициент на корелация може да приема стойности в интервала $-1.00 \leq r_{pb} \leq 1.00$. Положителните стойности на коефициента са свидетелство, че изпитаните лица, които са отговорили правилно на даден въпрос, имат относително високи тестови балове, а онези, които са отговорили неправилно – относително ниски. Това е и очакваният, логичен модел на връзката между индивидуалните отговори на даден въпрос и тестовия бал – предполага се, че по-подготвените (с по-висок тестов бал) са отговорили правилно на въпроса, а по-слабо подготвените – не. Този модел, който рефлектира в положителни (високи) стойности на коефициента, е свидетелство за добрите дискриминативни възможности на съответния въпрос.

• *Индексът на налучкване на верния отговор* отразява вероятността изпитваните да отговорят правилно на въпроса, без да имат необходимите знания. Този шанс обаче намалява при увеличаване на броя на алтернативите, а зависи и от умелото им съставяне. За премахване на влиянието на този фактор може да се въведе т. н. "поправка за налучкване", при която тестовият бал се коригира в зависимост от оценката на този индекс.

1.1.8. Статистики на алтернативните отговори

Това са група от показатели, които обикновено не се разглеждат като характеристики на въпросите, макар и да носят аналогична информация. Разликата е в това, че тези статистики се изчисляват за всеки алтернативен отговор, въз основа на броя на изпитаните лица, посочили съответната алтернатива.

За описание на алтернативните отговори най-често се използват следните статистики:

• *Дял на изпитаните, посочили съответния алтернативен отговор*. Изразява се чрез съотношението (като пропорция или в проценти) между изпитаните, посочили съответната алтернатива, и общия брой на изпитаните. В своята съвкупност, тези дялове формират честотното разпределение на всички изпитани между различните алтернативни отговори. Стойността на тази статистика, определена за ключовия отговор, се разглежда като индекс на трудността на целия въпрос.

• *Дял на изпитаните от силната/ слабата група, посочили съответния ал-*

тернативен отговор. Тези две статистики се изчисляват по същия начин, но нормирането е спрямо броя на изпитаните лица, принадлежащи към съответната група. Разликата между стойностите на тези две статистики, определени за ключовия отговор, се разглежда като индекс на дискриминативността на целия въпрос.

- *Точково-бисериален коефициент на корелация* между съответния алтернативен отговор и тестовия бал. Тук дихотомичната променлива, съдържаща отговорите на изпитаните лица, се кодира с 1, ако дадени лице е посочило съответния алтернативен отговор, и с 0, ако е посочили друг отговор. Стойността на тази статистика, определена за ключовия отговор, се разглежда като втори индекс на дискриминативността на целия въпрос.

Очевидно е, че статистиките на алтернативните отговори се определят по същия начин, както и тези на отделните въпроси. Това е така, защото алтернативите са натоварени със същите значения, имат същите характеристики, както и тестовите въпроси. Делът на посочилите съответната алтернатива има значение на нейна "трудност", разликата между дяловете на лицата от силната и слабата група или точково-бисериалният коефициент на корелация – на нейна "дискриминативност". Нещо повече, статистиките на една от алтернативите – ключовата, се разпростират върху целия въпрос. Поради това различното, което носи тази група от показатели, е свързано с дистракторите.

Статистиките на алтернативните отговори са важен източник на информация при апостериорното усъвършенстване на тестовите въпроси, особено при анализа на ефективността на дистракторите. Функцията на подвеждащите отговори, както показва тяхното наименование, е да “привлекат” към себе си вниманието на по-слабо подготвените лица, като ги отклонят от ключовия отговор. Разликата е в това, че статистиките на дистракторите следва да се интерпретират ако не по противоположен, то поне по различен начин от тези на ключовия отговор. Поради това един дистрактор е по-ефективен, ако е посочен като правилен от повече изпитани или ако има по-висока, отрицателна корелация с тестовия бал.

1.1.9. Предимства и недостатъци на Класическата теория

Като най-съществено преимущество на Класическата теория може да се отбележи това, че данните, които се получават в психологическите или образователните измервания чрез тестови процедури, обикновено удовлетворяват основните допускания в нейните модели. Поради това те обикновено са обозначавани като „меки” модели, които могат да бъдат приложени при различни по своите цели и характер измервания (Hambleton & Jones, 1993; Fan, 1998). Ако към това добавим достъпния статистически апарат, лесната интерпретация на тестовите статистики и понятното за широката аудитория значение на получения тестов бал, ще получим едно от възможните обяснения за широкото използване на Класическата теория при решаването на най-

разнообразни изследователски и практически задачи.

Друго предимство на Класическата теория е това, че анализите на тестовите характеристики могат да бъдат извършени въз основа на сравнително малки по обем представителни извадки. Това нейно качество е особено важно при пилотните изследвания, предназначени за оценка на качеството на конструирания тест, тъй като повишава тяхната ефективност.

Въпреки безспорните си качества Класическата тестова теория се характеризира и с някои съществени ограничения. На първо място трябва да посочим обстоятелството, че тестовите статистики и индексите на въпросите са зависими от извадката, чрез която са получени (*sample dependency*). Техните стойности не само се променят в различни извадки, но и придобиват смисъл единствено в контекста на съответната извадка или популация. Зависими от извадката са например дисперсията на действителния бал и надеждността, свързана с нея. Но, от друга страна, такава зависимост не се наблюдава при оценката на действителния бал на индивидуално равнище, нито при оценката на индексите на въпросите при основно τ-еквивалентните тестове, които са независими от популацията (Steyer, 2001).

Изследователите фокусират вниманието си върху зависимостта от извадката, чрез която са получени (*sample/ group dependency*), на двете основни статистики на тестовите въпроси – тяхната трудност и дискриминативна сила. Ако извадката е съставена от лица с по-високи знания и умения, оценките за трудността на въпросите биха били по-ниски и обратно – ако е съставена от лица с по-слаби знания и умения, трудността на въпросите би била по-висока. Индексът на дискриминативност на въпросите се влияе от друго качество на извадката – нейната хомогенност по отношение на измервания признак. По-високи стойности на индекса се получават при хетерогенни извадки (поради различията в силната и слабата група), докато хомогенните извадки биха довели до понижаване на неговите стойности. Тази особеност на двата индекса, която произтича пряко от методите за изчисляването им, ограничава тяхната пригодност като оценки на съответните характеристики на въпросите.

Трудността на отделните въпроси в тяхната съвкупност рефлектира върху трудността на теста, която, от своя страна, влияе пряко върху тестовите резултати. Така наблюдаваният тестов бал се оказва зависим от характеристиките на теста (*test dependency*) - той може да се понижи или повиши с изменение на трудността на теста. Изпитаните биха получили по-високи балове, ако се явят на тест, съставен от по-лесни въпроси, и по-ниски балове, ако тестът е съставен от по-трудни въпроси.

Наблюдаваният бал, следователно, може да се разглежда като една, при това неточна, оценка на действителния бал. Поради обстоятелството, че трудността на теста може да варира в зависимост от извадката, наблюдаваните балове, получени от различни тестове, не са пряко съпоставими (Hambleton, 2000).

Като съществен недостатък на Класическата теория се разглежда и нейното

безразличие към проблема за моделиране на връзката между (латентния) действителен бал и отговорите на изпитаните на отделните тестови въпроси. Така изследователят е лишен от инструментариум за предсказване на поведението на изпитваното лице спрямо даден конкретен въпрос (Steyer, 2001).

Класическата теория позволява определянето на една единствена стойност на надеждността на измерването, т.е. прави се едно нереалистично допускане за еднаква надеждност на измерването за всички изпитани лица, без оглед на техния действителен бал. Правени са наблюдения, според които оценяването е по-неточно в лявата (при слабите) и в дясната част на разпределението на тестовите балове скалата (при силните изпитани лица) и по-точно при изпитваните със средни възможности.

Като недостатък се разглежда и прекаленото фокусиране на теорията към стандартната грешка на измерване, която също е константна за целия тест, като се игнорират другите възможни фактори, влияещи върху тестовите резултати. Т. Доусън посочва три типа грешки, присъщи на всяко изследване, които могат да се отнесат и към психологическите измервания: (1) извадкова грешка (*sampling error*), (2) грешка на модела (*model specification error*), за каквато може да се приеме например изваждането на системните грешки от основното уравнение на CTT, и (3) случайна грешка на измерването (*measurement error*) (Dawson, 2003).

CTT разглежда грешката на измерване като случаен, несистемен компонент на действителния бал. Проблемът, според С. Даунинг, Т. Халадина и други автори, е в това, че някои променливи, които могат да бъдат потърсени сред елементите на измерването, представляват източник на системни грешки, които се добавят към действителния бал (Downing & Haladyna, 2006; Kline, 2005). Поради това грешката на измерването също може да се разглежда като величина с композитен характер. Освен случайния компонент, тя може да съдържа и втори адитивен компонент - системна грешка (*systematic error, bias*), т.е. $\varepsilon = \varepsilon_r + \varepsilon_s$. Това е компонент, който приема еднакви стойности при многократно измерване на един и същи обект. Поради системния си характер тези грешки корелират както с наблюдавания бал, така и със случайната грешки. Въпреки че източниците на системна вариация са изключени от основното уравнение на CTT, те неизменно влияят на надеждността и валидността на измерването. За да отразят това обстоятелство, С. Даунинг и Т. Халадина предлагат следното уравнение за дисперсията на наблюдавания бал, което представлява модификация на уравнение (20) (Downing & Haladyna, 2006):

$$\sigma_X^2 = \sigma_\tau^2 + \sigma_\varepsilon^2 + \sigma_s^2 + \sigma_{\varepsilon s} \qquad (37)$$

където:

σ_s^2 - дисперсия на системната грешка

$\sigma_{\varepsilon s}$ - ковариация между системната променлива и случайната грешка, отразя-

ваща връзката на тази променлива с наблюдавания бал и случайната грешка

За разлика от случайните грешки, системните грешки са предвидими като посока и размер (абсолютна стойност) и обикновено са пропорционални на действителната стойност. Ако техният източник бъде установен, той може да бъде контролиран, а системната грешка – ако не отстранена, то поне редуцирана.

Друга сериозна критика към CTT е начинът на формиране на суровия тестов бал (Steyer, 2001). Както беше отбелязано, прилага се практическото правило наблюдаваният бал се образува като сумарна скала на въпросите, без това правило да е аргументирано теоретично. То обаче е дискусионно и подлежи на обсъждане, тъй като вероятно са възможни и други начини за кумулиране и обобщаване на отговорите на отделните въпроси, например като претеглена сума, чрез средната им стойност и др.

1.2. Теория за отговор на тестов въпрос (IRT)

1.2.1. Обща характеристика

Появата и развитието на IRT, нейното налагане като основа на психологическите измервания, се разглежда от мнозина изследователи като „тиха революция" в оценяването (Embretson & Reise, 2000, стр. 13). Въпреки това емпиричните процедури на измерването в рамките на тази теория не се различават съществено от тези при Класическата теория. Най-често се разработва или се използва готов специализиран инструмент за измерване, съставен от множество въпроси, всеки от които е ориентиран към отделен аспект от съответната способност, представляваща изследователски интерес. Отговорите на и. л. се оценяват дихотомично – като правилен (на и. л. се приписва 1 точка) или като неправилен (на и. л. се приписват 0 точки). Необходимостта от този начин на скориране на отговорите предопределя във висока степен обстоятелството, че предпочитаният вид въпроси в инструментите за измерване са въпросите със структуриран отговор (обективни въпроси), чиято форма съответства на това изискване. От гледна точка на модела „Данни единичен стимул" (Coombs, 1964) дихотомичното скориране на отговорите отразява отношенията на доминиране между разноименните точки (на индивиди и на въпроси) на латентния континуум.

Двете теории следователно оценяват личностните черти посредством един и същи емпиричен механизъм, въз основа на отговорите на и. л. на тестовите въпроси. Базата от емпирични данни, върху която стъпват, също е една и съща: двумерна матрица от типа $A = C \times Q$ (виж уравнение 11), съдържаща дихотомичните оценки на отговорите на изпитаните лица. Съвършено различен е обаче начинът, по който двете теории използват регистрираните наблюдения за оценка латентните способности на индивидите и в крайна сметка за тяхното диференциране. При Класическата теория наблюдаваният бал се формира като сума от правилните отговори, по-точно като сума от

точките, приписани на отговорите на и. л. на различните въпроси, включени в теста (Steyer, 2001). Този бал (или някоя от неговите производни) служи за пряка оценка на действителния бал на индивида. IRT се фокусира най-напред върху отговорите на и. л. на всеки отделен въпрос и въз основа на тази информация, чрез съответния математически апарат, дава възможност за оценка на ненаблюдаемите индивидуални способности.

Онова, което решително отдалечава моделите на IRT от Класическата тория, може да бъде обобщено в следните няколко пункта. (1) IRT премества тежестта на анализа на по-ниско ниво – от тестовия бал, което е характерно за CTT, към отделния тестов въпрос[4]. (2) В по-широк план, IRT съществува под формата на различни модели и може да се разглежда по-скоро като обща теоретична концепция за обяснение на латентните чрез манифестираните променливи. Теорията включва разнообразни модели на връзката между изпълнението на тестовите въпроси от изпитаните лица и техните способности чрез прилагане на вероятностни подходи. Това прави измерването в рамките на IRT базирано на модел (*model-based measurement*) (Weiner et al., 2003). (3) Тестовите въпроси (чрез един от своите параметри) и изпитаните лица се разполагат на един и същи континуум и образуват смесено метрично пространство. (4) Една от най-интересните особености на моделите на IRT е прилагането на един принцип, който е по-скоро противоположен на споменатия по-горе втори принцип Жан-Пол Бензекри за разработването на модели, според който моделът трябва да пасва на данните, а не обратно. При калибрирането на теста се използва алгоритъм, който може да „изхвърли" част от данните (отделни въпроси или и. л.), които не съответстват на избрания модел.

1.2.2. Основни идеи и понятия

1.2.2.1. Латентни черти

Обект на всяко измерване в областта на психологията и образованието е една или друга характеристика на индивидите. Независимо дали интересът е насочен към тревожността, агресивността, локуса на контрол или към когнитивните способностите, тези характеристики споделят една обща особеност – те са ненаблюдаеми пряко. Изследователят може да съди за степента, в която даден индивид притежава една или друга характеристика, по нейните външни проявления, например по поведенческите прояви на тревожност и агресия или по бала на индивида от въпросници за измерване на тези характеристики. В областта на образованието такива външни белези са времето, което индивидът отделя за подготовка, бързината на усвояване на новия материал, оценките за постиженията в училище и др.

[4] Тази особеност на теорията намира отражение в нейното наименование.

Концепцията за латентните черти (*latent traits*) е въведена и популяризирана в науката от П. Лазасфелд във връзка с разработките му по латентно-структурния анализ (Lazarsfeld, 1960; Lazarsfeld & Henry, 1968; Lazarsfeld, 1969; Лазарсфельд, 1973). Терминът „латентен" в този контекст означава „скрит", „неявен", „който не може да бъде наблюдаван директно". Латентните черти не могат да бъдат измерени пряко така, като физичните свойства. Измерването им е винаги индиректно, чрез наблюдение на поведението на индивиди, изпълняващи релевантни задачи, или чрез отговорите на въпроси в различни психометрични инструменти. В този смисъл понятието за латентна черта е свързано с понятието за манифестирана черта (*manifested trait*) и придобива смисъл единствено чрез отношението си с него. Манифестираната черта е видим, достъпен за наблюдение израз на латентната. Наблюдаваното поведение се влияе от латентната черта, т.е. манифестираната променлива се разглежда като функция на латентната, тъй като между двете съществуват причинно-следствени отношения. Задачата на психометричните теории и модели е да се установи характера на тази функционална връзка, чрез което да се даде възможност за точна оценка на степента, в която индивидът притежава една или друга латентна черта. (Embretson & Reise, 2000; DeVellis, 2003).

В областта на образователното тестиране основните функционални типове тестове се обозначават като тестове за постижения (*achievement tests*) и тестовете за измерване на склонността за обучение (*aptitude tests*). При приложението на двата типа тестове обаче не може да се очертае някаква рязка граница и за преодоляване на тази терминологична несъгласуваност двата термина „все по-често се заменят с по-неутралния термин *ability* в наименованията на средствата за оценка на когнитивното поведение" (Анастази и Урбина, стр. 516-517). Поради това, а и поради огромното разнообразие от латентни черти, които могат да бъдат обект на измерване, в рамките на IRT като обобщаващо, родово понятие на знанията и уменията в различни предметни области се използва понятието „способност" (*ability*) (Baker, 2001, стр. 5). Тук и по-нататък в текста терминът „способност(и)" ще бъде използван именно в този смисъл, като синоним на термините „знания" „умения", „постижения" и „компетентности".

1.2.2.2. Характеристична крива на въпроса

Концепцията за характеристичната крива на тестовия въпрос (*item characteristic curve, ICC*) е основата, върху която се надграждат всички останали теоретични конструкти в моделите на IRT (Lord, 1980; Hambleton, Swaminathan & Rogers, 1991; Baker, 2001; DeVellis, 2003; Weiner et al., 2003). Поради това на този конструкт, както и на значението му в IRT, е необходимо да се отдели подобаващо внимание.

Подобно на психологическите черти, способностите на индивида в определена област могат да бъдат разгледани като латентна черта (или психологическа променлива). Тази латентна черта може да бъде представена като едномерно пространство

(континуум), а всеки индивид, който притежава такива способности – като точка на този континуум. Позицията на всеки индивид на континуума се определя от равнището (количеството) на неговите способности, което може да бъде изразено чрез определена числова стойност (*ability score*), обозначавана с Θ (тета). Тъй като в общия случай различните индивиди се различават по своите способности, техните точки биха били позиционирани на различни места на континуума, т.е. биха били асоциирани с различни числови стойности на Θ.

Съгласно втората теорема на П. Супес и Дж. Зинес (Суппес и Зинес, 1967), съвкупността от и. л., заедно с приписаните им числови стойности, отразяващи равнището на техните индивидуални способности, образуват скала, скала на компетентността (*ability scale*), която най-често се обозначава също с Θ. От теоретична гледна точка скалата на латентните способности е неограничена отляво и отдясно, поради което скаловите стойности на индивидите (индивидуалните Θ_i) могат да варират от отрицателна до положителна безкрайност. Съгласно модела „Данни единичен стимул" на К. Кумбс (Coombs, 1964), основната цел на психологическите измервания, включително и тези в образованието, може да бъде представена като построяване на скала на латентната черта, т. е. изпитаните лица да бъдат локализирани на континуума на способностите Θ, като им се припишат адекватни числови стойности Θ_i, отразяващи индивидуалното равнище на всяко изпитано лице.

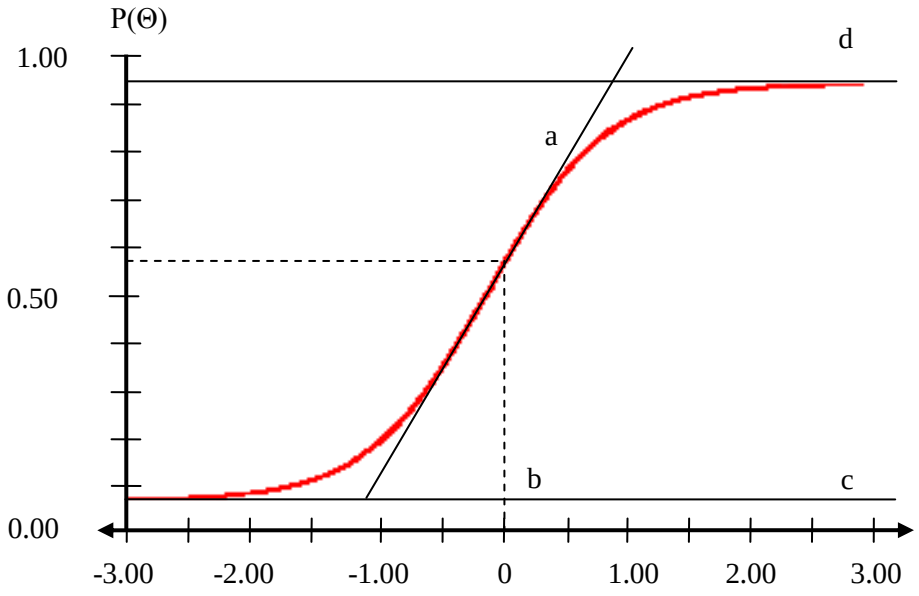
Беше отбелязано, че IRT фокусира вниманието си върху отделния въпрос и как всеки индивид е отговорил на него. При моделирането на този проблем в рамките на теорията се прилага вероятностният подход, което прави понятието „вероятност от правилен отговор" централно при анализа на връзката между индивидуалните равнища на латентната черта и отговорите на индивидите.

В общия случай, отделните индивиди се различават по своите способности (позиционирани са в различни точки на скалата Θ) и поради това биха отговорили правилно на даден въпрос с различна вероятност $P(\Theta)$, която може да варира в границите $0.00 \leq P(\Theta) \leq 1.00$. За разлика от Гутмановия модел, допускането в IRT е, че тази вероятност се изменя плавно, а не скокообразно. Тя би била сравнително по-висока при тези индивиди, които имат по-големи способности и съответно по-ниска при онези, които имат по-малки способности. IRT разглежда вероятността $P(\Theta)$ изпитаните лица да определят правилния отговор на даден въпрос като функция от техните способности. Графичното изображение на тази функция, представено на фигура 2, е характеристичната крива на въпроса, която има формата на плавна, монотонно нарастваща S-образна крива, неограничена отляво и отдясно.

Ординатите на характеристичната крива в левия край на континуума на способностите имат ниски стойности (лицата със слаби способности, локализирани в тази зона, биха отговорили правилно на въпроса с вероятност, близка до 0.00), издига се плавно нагоре (лицата със средни способности биха дали правилен отговор с вероят-

ност около 0.50), за да премине към десния край на скалата с ординати, близки до 1.00 (лицата с високи способности почти сигурно биха дали правилен отговор).

Фигура 2. Общ вид на характеристичната крива на въпрос



Тази форма на характеристичната крива съответства на допускането, че с нарастването на способностите нараства и вероятността от правилен отговор.

За описание на функционалната връзка между скалата на способностите и отговорите на съответния въпрос в по-голяма част от моделите на IRT се използва логистичната функция, предпочитана пред нормалната огива поради опростените изчислителни процедури (Lord, 1980; Embretson & Reise, 2000; Weiner et al., 2003; DeVellis, 2003).

Характеристичната функция на въпроса, в по-общ план, описва връзката между латентната и манифестираната променлива. В математически аспект това е връзката между скалата на способностите и вероятността от правилен отговор на даден въпрос. S-образната крива от фигура 2 се характеризира с 5 свойства: (1) наклон на кривата в средната ѝ част, (2) позиция спрямо хоризонталната ос (скалата на способностите), (3) долна хоризонтална асимптота, (4) горна хоризонтална асимптота и (5) симетричност. Тези няколко свойства са достатъчни за описанието на всяка характеристична крива, следователно и на всякакъв тип връзки между способностите на индивида и изпълнението му на тестовия въпрос.

1.2.2.3. Параметри на въпросите

Всеки въпрос в теста може да бъде описан чрез уникална характеристична крива. В рамките на IRT свойствата на кривата се определят като нейни параметри, които се атрибутират на съответния тестов въпрос.

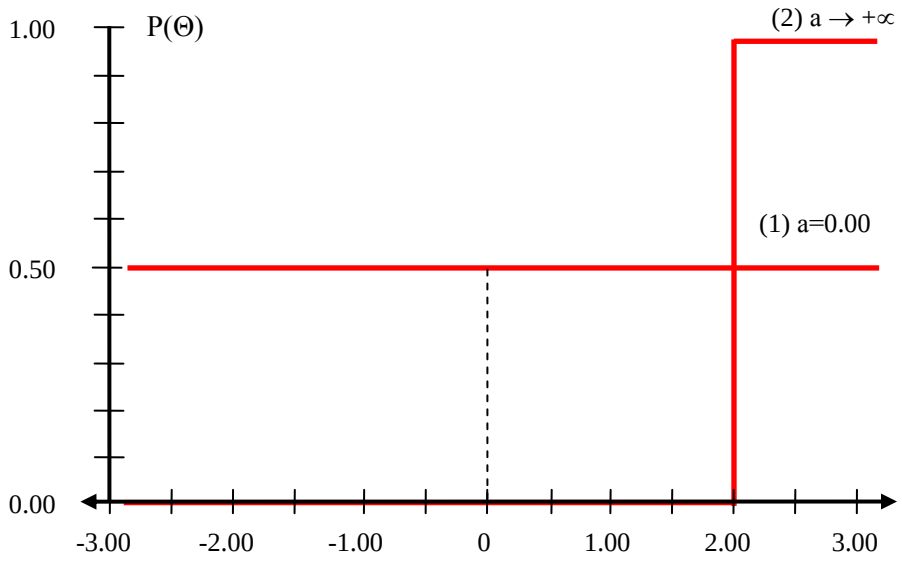
Параметър a – наклон/ дискриминативна сила на въпроса (*discrimination*). Най-общо определя формата на характеристичната крива, като от специален интерес е наклонът в средната ѝ част. Това свойство позволява на въпроса да разграничава индивидите, които имат по-ниски нива на способности от тези, които имат по-високи нива. Колкото по-стръмна е кривата в средната си част, толкова по-рязък е контрастът между вероятността за правилен отговор на лица с различни нива на способности. И обратно – колкото по-полегата е кривата, толкова по-слаба е диференциращата сила на въпроса, тъй като индивиди с различни нива на способности биха имали почти еднакъв шанс да отговорят правилно на този въпрос.

Теоретично стойностите на параметъра a могат да варират от $-\infty$ до $+\infty$, но при типичния въпрос те са в границите на $0.50 \leq a \leq 1.50$. При тези широки теоретични граници на изменение на параметъра е необходимо да се обърне внимание на два специални (гранични) случая на формата на характеристичната крива, която той задава. Докато при типичния въпрос формата на характеристичната крива е S-образна, при въпроси с нулева дискриминативна сила ($a = 0.00$) тя приема формата на хоризонтална линия, разположена на равнище $P(\Theta) = 0.50$, а при въпроси с максимална дискриминативност, клоняща към безкрайност – z-образна форма, характерна за айтем от детерминистичната Гутманова стъпкова функция, както е показано на следващата графика. При първия случай вероятността от правилен отговор на индивидите от всички равнища на способности Θ е еднаква, докато при втория случай вероятността от правилен отговор на индивиди, разположени вляво от $\Theta = 2.00$ (където е позициониран въпросът), е равна на 0.00, а на тези вдясно е равна на 1.00. Очевидно е, че примерният въпрос (2) с $a \to +\infty$ (или с $a \to -\infty$, при който хоризонталните сегменти биха имали противоположна ориентация) различава добре само онези индивиди, които се намират непосредствено в зоната около точката, в която е позициониран въпросът. Впрочем, това се отнася за всички въпроси с висока дискриминативна сила (Embretson & Reise, 2000; Baker, 2001; DeVellis, 2003).

Параметър b – позиция/ трудност на въпроса (*difficulty*). Определя мястото на характеристичната крива на континуума на способностите. Даден индивид i с равнище на компетентност Θ_i би имал по-голям шанс да отговори правилно на въпрос, чиято характеристична крива е разположена наляво от неговата точка (лесен въпрос), отколкото на въпрос с характеристична крива, позиционирана надясно от нея (труден въпрос). В по-общ план параметърът b описва в коя част на скалата на способностите функционира съответният въпрос. Параметърът може да приема стойности в интервала от $-\infty$ до $+\infty$, но обикновено се изменя в границите на $-3.00 \leq b \leq +3.00$. Трудните въпроси функционират в областта на по-високите стойности на Θ (т.е. в десния край на скалата), а лесните – в областта на по-ниските стойности (в левия ѝ край). Както е показано на фигура 2, всеки въпрос има скалова стойност за b (точка на континуума на способностите), която е вертикална проекция на инфлексната точка на характеристич-

ната крива, съответстваща на вероятност от правилен отговор $P(\Theta) = 0.50$.

Фигура 3. Характеристични криви на въпроси при $a = 0.00$ и $a \to +\infty$



Следователно трудността на въпросите и способностите на индивидите в теоретичната рамка на IRT са разположени на един и същи континуум и образуват смесено едномерно пространство (Coombs, 1964; Embretson & Reise, 2000; DeVellis, 2003; Downing, 2006).

Параметър c – долна асимптота/ налучкване на правилния отговор (*guessability, pseudo-guessing*). Показва вероятността от правилен отговор в случаите, в които способностите на индивидите са на ниско ниво. Изразява се чрез долната хоризонтална асимптота на логистичната крива, която определя вероятност от правилен отговор $P(\Theta) > 0.00$. Ако стойността на този параметър е по-висока от 0.10, тя се интерпретира като значима. Параметърът отразява силната мотивация на някои индивиди за постигане на високи постижения, дори и изпълнението на теста да е предшествано от недостатъчни усилия за подготовка. Тази мотивация намира израз в прилагането на стратегия на безусловно посочване на една от алтернативите, при която изпитваните лица се ориентират към ключовия отговор не въз основа на знания и умения, а по външните белези на тестовия въпрос (структура, дължина на дистракторите, езикови формулировки и др.) или чрез случаен избор (Embretson & Reise, 2000).

Параметър d – горна асимптота/ невнимателност (*carelessness*). Показва вероятността от правилен отговор в случаите, в които способностите на индивидите са на високо ниво. Изразява се чрез горната хоризонтална асимптота на логистичната крива, която определя вероятност от правилен отговор $P(\Theta) < 1.00$. Параметърът отразява влиянието на комплекс от фактори, които възпрепятстват лицата с по-високи способ-

ности да посочат коректния отговор. Това може да бъде по-слабата на мотивация на някои индивиди за постигане на високи постижения, дори и да са достатъчно добре подготвени за изпита. Тя се изразява или в отбягване на какъвто и да е отговор, ако индивидът е неуверен, или в посочване на неправилен отговор поради по-слаба концентрация на вниманието, неправилно или неточно осмисляне на задачата, поставена чрез въпроса, небрежност, нехайство за крайния резултат и дори лекомислие и безотговорност. В някои случаи посочването на некоректен отговор може да бъде поради техническа грешка или заблуждение (*mistake-ability*), както и поради силната привлекателност на някой от дистракторите, твърде близък до коректния отговор (*attractive distractor parameter*) (Barton & Lord, 1981; Linacre, 2004).

Параметър e - асиметрия (*asymmetry*). Моделите на IRT, които можем да определим като „стандартни", се базират на такива криви на въпросите, които се характеризират с точкова симетрия. Допускането за симетричност на кривата обаче не винаги съответства на емпиричните данни. Наблюденията, направени от Ф. Самеджима върху данни от различни психологически изследвания, разкриват системни несъответствия между трудността на въпросите и подредбата на индивидуалните способности Θ, водещи до грешки в техните оценки (Samejima,1995; 1997; 2000). Авторката въвежда семейство от логистични експоненциални модели, в които предлага асиметрична форма на характеристичните криви на въпросите, чрез която забелязаните несъответствия могат да бъдат преодолени. Разбира се, асиметричността на характеристичната крива може да бъде наблюдавана при стойност на долната асимптота (параметър) c = 0.00 и на горна асимптота (параметър) d = 1.00. Асиметрията на характеристичната крива (отрицателна или положителна) се изразява в изменение на нейния наклон в някои сегменти, което води до изменение на вероятността от правилен отговор в сравнение със съответната симетрична крива. При отрицателна асиметрия тази вероятност се увеличава при (част от) лицата с по-високи способности, а при положителна асиметрия – намалява при (част от) лицата с по-ниски способности. Това дава основание на някои автори да разглеждат този параметър като наказателен/ награждаващ (Bazan, Bolfarine & Branco, 2004).

Параметрите на тестовите въпроси са независими един от друг. Въпрос с дадена трудност може да има различни стойности за параметрите a, c и d и обратно – въпрос с фиксирана дискриминативност може да има различни позиции на континуума на способностите. В общия случай, формата на характеристичната крива на всеки тестов въпрос представлява уникална комбинация от стойностите на нейните параметри.

1.2.2.4. Характеристична крива на теста

Подобно на CTT, и IRT борави с понятията "суров/ наблюдаван тестов бал" (*raw test score*) и "действителен бал" *(true score*). В рамките на модерната теория те имат

аналогични значения – съответно на сума от дихотомично кодираните отговори на изпитаното лице на въпросите от теста и на средна стойност на множеството от наблюдаваните тестови балове на това лице, получени при независими тестирания.

Различен обаче е начинът за определяне на действителен бал, който се използва в IRT. Съгласно подхода, предложен от Д. Лоули, формулата за определяне на действителния бал е следната (Lawley, 1943, цит. по Baker, 2001):

$$TS_i = \sum_{j=1}^{k} P_j(\Theta_i) \qquad (38)$$

където:

TS_i - действителен бал на изпитаните лица с равнище на компетентност Θ_j

j - пореден номер на тестов въпрос

k - брой на въпросите в теста

$P_j(\Theta_i)$ - вероятност от правилен отговор на j-тия въпрос от изпитаните лица с равнище на способности Θ_i

$P_j(\Theta_i)$ се определя в зависимост от конкретния модел на характеристичната крива на дадения въпрос. Както се вижда, съгласно подхода на Д. Лоули действителният бал за дадено равнище на компетентност се определя като сума от вероятностите за правилен отговор на всички въпроси в теста. Тъй като вероятността за даден въпрос се изменя в границите $0.00 \le P_j(\Theta) \le 1.00$, то действителният бал може да приеме стойности в границите $0.00 \le TS \le k$.

Кривата, която отразява функционалната връзка между скалата на способностите Θ и действителния бал TS, е характеристичната крива на теста. Тя има форма на монотонно нарастваща функция, която не се задава чрез конкретен математически израз, не се описва чрез параметри и поради това нейната форма е конкретна за всеки тест. Обикновено тя е S-образна, подобна на характеристичната крива на въпросите. По подобен начин формата на характеристичната крива не зависи от разпределението на изпитаните лица на скалата на способностите. При прилагане на 1PL и 2PL модели, левият й край клони неограничено към 0.00, когато равнището на компетентност клони към $-\infty$, а десният й край – към k, когато равнището на компетентност клони към $+\infty$. При прилагане на 3PL модел, левият край на кривата се приближава неограничено към сумата от стойностите на параметрите на налучкване на въпросите в теста $\sum_{j=1}^{k} c_j$, а десният й край - към k (Baker, 2001).

В по-общ план характеристичната крива на теста е периферен конструкт в понятийната система на IRT. Тя играе важна роля през етапа на интерпретиране и представяне на резултатите от теста, като служи за преобразуване на скалата на способностите в скала, която е свързана с тестовия бал и поради това е лесно разбираема за

потребителите на тестовите резултати. Но тя носи важна информация и за съставителя на теста. Разположението на кривата на хоризонталната ос на равнище $k/2$ определя скаловата стойност на теста, т.е. общата му трудност. Нейният наклон отразява начина, по който са свързани действителния бал и скалата на способностите (Lord, 1980; Embretson & Reise, 2000; Baker, 2001; DeVellis, 2003).

1.2.2.5. Прецизност (стандартна грешка) на оценката

Крайната цел на всяка процедура на скалиране е да се определи скаловата стойност на измерваемия обект по отношение на измереното свойство. За IRT задачата е да се построи скала на изпитаните лица, т.е. всяко от тях да бъде локализирано на континуума на способностите. Равнището на компетентност на индивида Θ_i, поточно действителната стойност на този личностов параметър, е неизвестна и от теоретична гледна точка остава такава. Както бе отбелязано при анализа на действителния бал TS, при многократни, независими тестирания на едно лице може да се получи серия от различаващи се наблюдавани тестови балове. По подобен начин, ако при всяко тестиране се прави оценка $\hat{\Theta}_i$ на действителната компетентност Θ_i на това лице, може да се получи серия от различаващи се оценки. Това означава, че всеки опит за оценка на личностовия параметър е асоцииран с определена грешка и нейната дисперсия е пряко свързана с прецизността на измерването (Embretson & Reise, 2000).

Като мярка за определяне на отклоненията на оценките $\hat{\Theta}_i$ от стойността на търсения личностов параметър Θ_i се използва стандартната грешка на оценката (*standard error of estimation*). Колкото по-малка е нейната стойност, колкото по-малка е вариативността на оценките, толкова по-прецизно е измерването на неизвестния параметър.

За оценка на действителните стойности на Θ в IRT най-често се използва методът на максималното правдоподобие (*maximum likelihood*). Това е итеративна процедура за повишаване на прецизността в оценката на личностовия параметър, основана на съпоставяне на актуалните стойности (1/ 0) на отговорите на дадено изпитано лице на въпросите от теста и асоциираните с тях вероятности за правилен отговор на това лице (Lord, 1980). Стандартната грешка при еднопараметричния модел се определя по следната формула (Baker, 2001):

$$SE(\hat{\Theta}) = \frac{1}{\sqrt{\sum_{j=1}^{k} P_j(\hat{\Theta})\, Q_j(\hat{\Theta})}} \qquad (39)$$

където:

k - брой на въпросите

$P_j(\hat{\Theta})$ - вероятност от правилен отговор на j-тия въпрос

$Q_j(\hat{\Theta}) = 1 - P_j(\hat{\Theta})$ - вероятност от неправилен отговор на *j*-тия въпрос

Подобно на CTT, и тук обемът на теста е в пряка връзка с прецизността на оценката. Прецизността нараства с увеличаване на броя на въпросите.

1.2.2.6. Информационна функция на въпроса/ теста

Оценките на неизвестния параметър Θ са единствените носители на информация за неговата стойност. Колкото по-близка е една оценка $\hat{\Theta}_i$ до параметъра Θ_i, толкова повече информация носи тя. Поради това отклоненията на оценките от действителната стойност на личностовия параметър, т.е. прецизността на измерването, е от ключово значение за качеството на тестовата процедура.

Концепцията за информативността на измерването в областта на статистиката, възприета след това и в психометрията, е една от многото плодотворни идеи на Р. Фишер. Той дефинира информативността като реципрочна на прецизността, с която е оценен даден параметър (Baker, 2001). И тъй като за мярка на прецизността се използва дисперсията на оценките около неизвестния личностов параметър, информативността може да се представи чрез следния общ математически израз:

$$I(\hat{\Theta}) = \frac{1}{Var(\hat{\Theta})} \qquad (40)$$

Очевидно е, че понятието за прецизност/ информативност в IRT има концептуална връзка с понятието за надеждност в CTT. Но, подобно на други конструкти в модерната теория, информативността на измерването може да бъде оценена за всяко конкретно равнище Θ_i на скалата на компетентността. В допълнение IRT, която е фокусирана върху отделните въпроси, предлага методи за оценка на прецизността/ информативността на измерването на този параметър на равнище тестов въпрос, а след това и на равнище тест.

Всеки тестов въпрос носи определена, макар и малка информация за способностите на отделния индивид. Кривата, която отразява функционалната връзка между скалата на компетентността Θ и количеството информация, определено за всяка нейна точка въз основа на даден въпрос, е информационната функция (*item information function*) на този въпрос (Lord, 1980; Embretson & Reise, 2000). И тъй като стандартната грешка на оценката SE е втори корен от дисперсията, то информационната функция на въпроса може да се определи, като в знаменателя на уравнение (40) се вземе квадратът на стандартната грешка от уравнение (39). По този начин бихме получили желаната мярка за информативността на измерването във всяка точка на скалата Θ. Математическата формула за определяне на информационната функция на въпросите зависи

от модела на IRT, въз основа на който са определени характеристичните им криви. В рамките на еднопараметричния модел информационната функция на въпроса се определя чрез следното уравнение:

$$I_j(\Theta_i) = P_j(\Theta_i) Q_j(\Theta_i) \quad (41)$$

където:

$I_j(\Theta_i)$ - стойност на информационната функция на j-тия въпрос на равнище на компетентност Θ_i

$P_j(\Theta_i)$ - вероятност от правилен отговор на същия въпрос от индивид с равнище на компетентност Θ_i

$Q_j(\Theta_i) = 1 - P_j(\Theta_i)$ - вероятност от неправилен отговор на същия въпрос от индивид с равнище на компетентност Θ_i

При 2PL модел, в горното уравнение се включва още един елемент – дискриминативната сила на въпроса:

$$I_j(\Theta_i) = a_j^2 P_j(\Theta_i) Q_j(\Theta_i) \quad (42)$$

където:

a_j - дискриминативен параметър на j-тия въпрос

Очевидно е, че при 2PL модел пряко влияние върху прецизността/ информативността на въпроса оказва и стойността на дискриминативния параметър.

Логично е при определянето на информационната функция в рамките на 3PL модел да бъде включен и параметърът за налучкване c.

$$I_j(\Theta_i) = a_j^2 \left[\frac{Q_j(\Theta_i)}{P_j(\Theta_i)} \right] \left[\frac{P_j(\Theta_i) - c_j^2}{(1 - c_j^2)} \right] \quad (43)$$

Както беше отбелязано, характеристичната крива на въпроса при 3PL модел не притежава характеристиките на логистична крива. Поради това и формулата на информационната функция се отличава с подчертана сложност. Формата на информационната крива е сходна с тези при предходните модели, ни поради присъствието на параметъра c обикновено е по-ниско разположена. Това означава, че в рамките на 3PL, в който се отчита влиянието, което стратегията на налучкване оказва върху резултатите от теста, информативният потенциал на въпросите е по-слаб, отколкото в останалите два модела.

Информационната функция на въпроса разкрива с каква прецизност е оценена стойността на неизвестния параметър в една или друга точка на Θ. Тъй като отделните

въпроси отразяват различни аспекти на латентната способност, не всички я представят и следователно измерват еднакво добре. Поради това всеки въпрос, в общия случай, се характеризира с конкретна, уникална информационна крива. Обикновено тя има изпъкнала, (приблизително) симетрична форма с един локален връх, разположен над онази точка на Θ, която съвпада с трудността на съответния въпрос, и с краища, чиито ординати намаляват в знак на намаляващата прецизност на измерването. Поради тази особеност оценките на различните равнища на компетентност не се характеризират с еднаква прецизност. И ако тези оценки са направени чрез използването само на един тестов въпрос, количеството информация в различните точки на Θ е твърде малко, дори и областта около b. Това е едно от основанията в теста да включва поголям обем от въпроси.

Тъй като тестът представлява съвкупност от въпроси, всеки от които се характеризира с определена информационна функция, е възможно да се определи информационната функция на целия тест. Тя представлява зависимостта между скалата на компетентността Θ и количеството информация, определена за всяка нейна точка въз основа на всички въпроси в теста. Поради допускането за локална независимост на отговорите, количеството информация за целия тест в дадена точка от латентния континуум е адитивно. Операционално то се дефинира като сума от количеството информация, която предоставя всеки въпрос в същата точка (Lord, 1980; Embretson & Reise, 2000).

$$I(\Theta_i) = \sum_{j=1}^{k} I_j(\Theta_i) \qquad (44)$$

където:

$I_j(\Theta_i)$ - количество информация от j-тия въпрос на равнище на компетентност Θ_i

Поради специфичния начин на определяне на стойностите на тестовата информационна функция, посочен по-горе, в хода на тестовия анализ най-напред се определят информационните функции на отделните въпроси. Ако всички те се характеризират с една и съща форма и позиция на скалата на способностите, тестовата информационна функция би имала изпъкнала симетрична форма, подобна на тази на информационните функции на отделните въпроси, с намаляващи ординати в двата и края. Общото й ниво обаче би било много по-високо, защото в нея е кумулирано цялото количество информация, предоставено от всеки един от тях. Очевидно е, че прецизността на измерването на тестово равнище може да се подобри както с увеличаване на броя на въпросите, така и с усъвършенстване на тяхното качество, т.е. с повишаване на собствената им информативност (Baker, 2001).

1.2.2.7. Оценка на параметрите на теста

Най-важната задача при психологическото оценяване, неговата, така да се каже, „свръхзадача", е да се построи скала на изпитваните лица. Това означава да се определи позицията на всяко едно от тях на континуума на латентната променлива, т.е. да намерят индивидуалните числови стойности на параметъра Θ. Заедно с това е необходимо да се определят числовите стойности на параметрите a, b, c, d, и e на тестовите въпроси, в зависимост от избрания модел на IRT (Lord, 1980; Baker, 2001; DeVellis, 2003). Тази процедура, позната като „калибриране на теста", в частност може да бъде разгледана като изграждане на метрично пространство.

При измерване на разстоянието между две точки x и y, лежащи на една права, това разстояние се определя като $|x - y|$, абсолютната стойност на разликите между техните координати. Това правило е изведено на абстрактно равнище в математическата концепция за метрично пространство.

Основно в тази концепция е понятието за непразно множество M, което се състои от n елемента ($n \geq 3$), между всеки два от които е установено някакво разстояние. В този смисъл под метрично пространство (*metric space*) се разбира двойката (M, d), в която d: $M \times M \rightarrow R$ е функция (метрика или функция на разстоянието) върху M, чрез която на всяка двойка елементи $x, y \in M$ се присвоява реално число $d(x, y)$. Функцията d следва да удовлетворява следните аксиоми за всяко x, y и $z \in M$:

(1) $d(x,y) \geq 0$ - неотрицателност (45)

(2) $d(x,y) = d(y, x)$ - симетричност (46)

(3) $d(x,y) \leq d(x,z) + d(y,z)$ - неравенство на триъгълника (47)

(4) ако $x = y$, то $d(x,y) = 0$ - идентичност (48)

(5) ако $d(x,y) = 0$, то $x = y$ - идентичност (49)

Ако са удовлетворени всички горни изисквания (45 - 49), тогава d е мярка за разстоянието или метрика на множеството от елементи на M. Двойката (X, d) се определя като метрично пространство. Общият вид на формулата за определяне на разстоянията в едно метрично пространство е следната:

$$d(i,j) = \left[\sum_{k=1}^{m} \left(a_{ik} - a_{jk} \right)^{l} \right]^{\frac{1}{l}} \quad (50)$$

където:

$d(i, j)$ - разстояние между точките i и j

m - брой координати на точките (брой оси на пространството)

a_{ik}, a_{jk} - проекции (координати) на точките i и j на оста k

l - показател, който определя типа на метриката

При $l = 1$ метриката се определя като абсолютна или линейна (*Manhattan*, *city-*

block).

При $l = 2$ това е позната Евклидова метрика (*Euclidean distance*).

При $l > 2$ метриката е известна като дистанция (l -метрика) на Минковски.

Процедурата на калибриране на теста е изграждане на метрично пространство, което може да бъда определено като смесено, съгласно модела „данни единичен стимул" на К. Кумбс (Coombs, 1964), защото като точки в пространството са представени както изпитаните лица, така и тестовите въпроси, между всеки две от които се определя съответната числова стойност $d(x, y)$ в метриката на латентната променлива. За $d(x, y)$ може да се мисли като за разстояние между съответните две разноименни точки.

Техниката за калибриране на теста, разработена от А. Бирнбаум (Birnbaum, 1968), е двустъпкова итеративна процедура, основана на метода на максималното правдоподобие. На първата стъпка се оценяват параметрите на тестовите въпроси, а на втората – личностовите параметри Θ_i, отразяващи равнищата на компетентност на изпитаните лица. На всяка стъпка процедурата за оценяване на съответните параметри се повтаря дотогава, докато се достигне до техните стабилни, непроменящи се стойности, които се различават минимално от стойностите, получени при предходната итерация. Характерно за процедурата на Бирнбаум е това, че оценката на параметрите на въпросите, а след това и на изпитаните лица се извършва едновременно (в рамките на една и съща процедура), въз основа на едни и същи дихотомично скорирани тестови резултати, макар че и въпросите, и изпитаните се оценяват последователно, един по един (Embretson & Reise, 2000; Baker, 2001).

Като съществен недостатък на тази процедура се посочва, че тя не води до еднозначно определяне на метриката на скалата на компетентността. Нека да върнем към понятието „метрика". Една функция d се нарича метрика, ако чрез нея на всяка наредена двойка (x, y) от елементите x и y на множеството M се съпоставя някакво реалното число (разстояние) $d(x, y)$ при спазване на горните пет условия (45 - 49). При процедурата на А. Бирнбаум на всяка двойка елементи могат да бъдат присвоени различни, макар и подходящи стойности. Като резултат средата на скалата и единицата мярка не са фиксирани и също могат да приемат различни стойности. Поради това е необходимо да се направят още допълнителни процедури за фиксиране на скалата чрез конвенционални техники за определяне на нейната среда и на единицата за измерване (Baker, 2001).

Например в алгоритъма за калибриране, използван в моделите на Раш, средната стойност на скалата се определя като средна стойност на оценките на трудността на въпросите b, т.е. скалата се центрира спрямо онзи интервал от континуума Θ, в който са локализирани характеристичните криви на въпросите. По-нататък скалата на трудността се трансформира линейно чрез добавяне на средната трудност като отрицателна константа към оценката на трудността на всеки въпрос b_j, така че средната трудност (средата на скалата) да бъде сведена до 0.00. Тъй като дискриминативният

параметър в този модел е с фиксирана стойност (обикновено a = 1.00), единицата на измерване на компетентността е фиксирана на същата стойност (Wright & Mead, 1976; Wright, 1999; Wrigt & Stone, 1999).

Въпреки това като краен резултат от тази процедура метриката на изграденото метрично пространство зависи от (1) специфичните характеристики на въпросите, изграждащи теста и (2) специфичните равнища на компетентност, с която се характеризират изпитаните лица. Поради това „не е възможно да се направят оценки на изпитаните лица и на параметрите на въпросите, които да бъдат с метриката на латентната променлива" (Baker, 2001, стр. 136).

1.2.3. Основни допускания в IRT

Теорията за отговор на тестов въпрос е по-„строга" и по-добре концептуализирана от КТТ, включително и по отношение на нейните основни допускания. Въпреки това различните автори посочват различен брой основни допускания. Някои автори определят като такива само допускането за размерността на латентната структура и за математическата форма на характеристичната крива на въпроса (Hambleton & Jones, 1993; Nandakumar, Yu, Li & Stout, 1998). Други причисляват към допусканията независимостта на латентната способност от съдържанието на теста, едномерността на теста, както и възможността за оценка на способността въз основа на различни множества от въпроси (Wiberg, 2004). Трети говорят за монотонност, размерност и локална независимост (Nandakumar & Ackerman, 2004). Р. Хамбълтън и колеги определят като основни допускания едномерността, еднаквата дискриминативност на въпросите и възможността за налучкване на правилния отговор (Hambleton, Swaminathan & Rogers, 1991). Трябва да се отбележи обаче, че тези различия се дължат не толкова на концептуална неяснота, колкото на това за кои модели на IRT се отнасят съответните допускания.

Все пак може да се очертае един минимален набор от най-често посочвани основни допускания, които да бъдат свързани с по-голяма част от моделите на IRT като тяхна обща теоретична рамка.

(1) Отговорите на индивидите на тестовите въпроси могат да бъдат предсказани (обяснени) чрез наличието на съвкупност от фактори, които представляват характерни черти (*traits*), скрити черти (*latent traits*) или способности (*abilities*) (Hambleton, Swaminathan & Rogers, 1991). Чрез това твърдение се заявява, че между това, което изследователят може да наблюдава и измерва пряко, и това, от което действително се интересува, но не може да измерва пряко, съществува някакъв тип каузална връзка.

(2) Връзката между съвкупността от фактори (ненаблюдаеми латентни черти) и отговорите на въпросите (наблюдаваните манифестирани променливи) може да бъде моделирана чрез някакъв тип монотонно нарастваща функция - характеристична фун-

кция на въпроса (*Item characteristic function, ICF*) или характеристична крива на въпроса (*Item characteristic curve, ICC*) (Hambleton, Swaminathan & Rogers, 1991; Hambleton & Jones, 1993; Harvey, 2003; Weiner et al., 2003; Kline,2005). Функционална връзка има определена математическа форма и познаването на характера на тази връзка дава възможност за получаване на информация за латентните черти по информацията, получена за/от наблюдаваните променливи. Най-общо, с нарастване на равнището на латентната променлива нараства вероятността от посочване на ключовия отговор.

(3) Всеки модел на IRT допуска наличието на един единствен вид на функционалната връзка между латентните и наблюдаваните променливи. (Downing & Haladyna, 2006). Това предполага, че характеристичната крива има точно определена форма. Следствие от това допускане е, че характеристичните криви всички въпроси в теста имат една и съща обща форма.

(4) Нормалност на разпределението на латентната променлива (*ability variable*)

Беше отбелязано, че основното статистическо допускане в моделирането на академичните постижения, а и на други променливи, свързани с човешкото поведение, е допускането за нормалност на разпределението на латентната променлива (Kline, 1998; Weiner et al., 2003; Holland & Hoskens, 2003; Kline, 2005) или на наблюдаваните балове (Bazan, Bolfarine & Branco, 2004). За разлика от Класическата теория, то е дефинитивно при IRT (Wiberg, 2004; Downing & Haladyna, 2006).

Съществото тук е, че процедурите за изграждане на характеристичната крива на въпроса (т.е. оценяване на параметрите на съответната функция) обикновено са базирани на алгоритъма на максималното правдоподобие (*Maximum likelihood estimation*, ML или *Marginal maximum likelihood estimation*, MML), който почива на няколко основни допускания, най-важното сред които е за нормалност на латентната променлива Θ в популацията със средна 0.00 и стандартно отклонение 1.00 (Breckler, 1990; Sočan, 2000; Brennan, 2001; Baker, 2001; Reynolds & Kamphaus, 2003; Kline, 2005).

При тази процедура нивата на значимост (*p-levels*), асоциирани с оценените параметри, също са базирани на нормалното разпределение (Kline, 2005). Нормалността на латентната променлива е основа за постояване на доверителни интервали около оценките на различните нива на способност $\hat{\Theta}$ (Hambleton, Swaminathan & Rogers, 1991). Ако Θ не е нормално разпределена, това води до неточно оценяване на параметрите на въпросите, въпреки че, според някои изследователи, оценките на параметрите не се влияят силно от умерени отклонения на Θ от нормалността (Embretson & Reise, 2000; Woods, 2006).

В предходната част бяха посочени редица статистически методи, основани на нормалното разпределение, които се използват при анализа на данни от психологически изследвания, включително и на академичните постижения. Тук ще обърнем внимание на факторния анализ, който обикновено се използва за оценка на размерността на скритото пространство на психологическите променливи. Той предполага латентни-

те променливи да бъдат континуални и нормално разпределени, а връзките между тях – линейни. Нарушенията на тези изисквания водят до подценяване на факторните тегла и надценяване на броя на латентните променливи (Embretson & Reise, 2000).

Отчитайки обвързаността на стандартните модели на IRT с допускането за нормално разпределение на моделираните когнитивни способности и други психични черти, някои автори поставят под въпрос доколко адекватни са тези модели (Micceri, 1989; Samejima, 1997). Ф. Самеджима обсъжда идеята за въвеждане на несиметрични по форма характеристични криви на въпросите, а Х. Базан и сътрудници разработват нови психометрични модели, които боравят с асиметрични разпределения. Авторите предлагат фамилия от асиметрични модели на IRT (*Skew-Normal IRT family)* за моделиране на индивидуалните способности (Bazan, Bolfarine & Branco, 2004). В основата им лежи идеята за кумулативно асиметрично-нормално разпределение на характеристичната крива на въпроса (в която асиметричността се задава от параметъра λ) и асиметрично-нормално разпределение на латентната променлива. Това разпределение генерализира стандартното нормално разпределение, в което към познатите два параметъра се добавя трети параметър k, „отговарящ" за степента на асиметричност.

В този интересен модел разпределението на латентната променлива U_i, съответстваща на i-тия индивид, може да бъде представено като:

$$U_i \sim SN(\mu, \sigma^2, k),\ i = 1, 2, \ldots, n, \qquad (51)$$

което представлява асиметрично-нормално (*skew-normal*) разпределение с централна тенденция $-\infty < \mu < +\infty$, разсейване $\sigma^2 > 0$ и параметър за асиметрия $-\infty < k < +\infty$. При $k = 0.00$ разпределението е строго симетрично.

(5) Едномерност на латентното пространство

Това допускане се изразява в това, че отговорите на и. л. на въпросите в теста са обусловени от една единствена латентна черта (фактор). Взаимовръзката (ковариацията) между отделните въпроси може да бъде обяснена само с тази черта (Hambleton, Swaminathan & Rogers, 1991; Hambleton & Jones, 1993; Harris, 1993; Fan, 1998). Това допускане може да бъде изразено по няколко начина, например, като едномерност на латентната структура, лежаща в основата на теста. Всички въпроси в теста измерват една единствена латентна променлива. Поради наличието на голям брой доказателства, че това допускане е твърде силно, някои автори говорят за „стриктна" (т. е. абсолютна, безусловна) едномерност и „основна" (*essential*) едномерност, при която се допуска многомерност, но с наличието на един доминантен фактор (Kingston, Leary & Wightman, 1985; Stout, 1987; Nandakumar, 1993; Fan, 1998; Harvey, 2003, Holland & Hoskens, 2003; Downing & Haladyna, 2006). Терминът „основна едномерност" е въведен от У. Стаут, който предлага специален математически модел за оценка на доминантността на дадена димензия (Stout, 1987).

Размерността на латентната структура се определя не само от конкретната способност (или способности), която е обект на измерване. В свое изследване М. Рекасе показва, че тя може да бъде функция и на самите тестови въпроси, на методите на преподаване, дори и на изпитваните лица, в зависимост от тяхното психично състояние, например равнището на тревожност в момент на изпита (Reckase, 1990*)*.

(6) Локална/ условна (*local/ conditional*) независимост на отговорите на и. л. на въпросите в теста

Локалната независимост е едно от основните допускания на IRT, тясно свързано с изискването за едномерност. Съгласно това допускане, връзката между айтемите се дължи единствено на (съответно може да бъде обяснена чрез) влиянието на латентната променлива. Ако това влияние бъде отстранено (т.е. при определена константна стойност на тази променлива), отговорите на и. л. на всеки два въпроса от теста са статистически независими, т. е. остатъчната ковариация между въпросите е равна на нула (Hambleton, Swaminathan & Rogers, 1991; Embretson & Reise, 2000; Kline, 2005; Downing & Haladyna, 2006). С други думи, локалната независимост предполага пълна липса на корелация (независимост) между отговорите на група (клас) и. л., разположени на една и съща позиция на континуума на латентната променлива, т.е. характеризиращи се с едно и също ниво на способност Θ.

Това допускане е обвързано с допускането за едномерност на латентното пространство, т.е. с едномерния модел на IRT. Поради това, ако в основата на теста лежат няколко променливи, това допускане следва да се разглежда като нарушено. Ф. Лорд и други автори обаче предлагат по-разширено тълкуване на изискването за локална независимост, което не нарушава неговия смисъл. Ако размерността на избрания модел на IRT е същата, каквато е размерността на данните, то допускането за локална независимост може да се разглежда като удовлетворено (Kingston, Leary & Wightman, 1985; Lord & Novick, 1974).

Част от посочените по-горе допускания не са валидни за всички модели на IRT. Все пак може да се посочи една по-устойчива схема, един най-често прилаган модел, който може да бъде определен като „стандартен" – това е едномерен, базиран на нормално разпределение на латентната променлива, логистичен модел.

Освен общите допускания, в рамките на отделните модели IRT се правят и редица допълнителни допускания, които определят спецификата на съответния модел. Такива са например допускането за равенство на дискриминационните индекси (Hambleton, Swaminathan & Rogers, 1991), което е ключово за еднопараметричния модел; за липса на налучкване на ключовия отговор (ibid.), което е специфично за едно- и двупараметричния модел; за логистичния вид на функцията – при логистичните модели и т. н.

Основната разграничителна линия между различните модели на IRT минава през броя и вида на характеристиките (параметрите) на въпросите, които обуславят

изпълнението им от и. л. (Hambleton, Swaminathan & Rogers, 1991; Downing & Haladyna, 2006).

1.2.4. Модели на IRT

Теорията за отговор на тестов въпрос съществува под формата на семейство от теоретични модели, предназначени за описание на различни типове данни. Обсъждайки възможните връзки между понятията „теория" и „модел", П. Супес отбелязва, че моделът е теоретична структура, представляваща „една възможна реализация" на дадена теория, в която „всички валидни твърдения на теорията са удовлетворени" (Suppes, 1962, стр. 252). Моделите на IRT, функциониращи в една обща концептуална рамка, съответстват напълно на тази формулировка.

Моделите могат да бъдат класифицирани въз основа на различни диференциални признаци, най-важни сред които са следните.

(1) Според размерността на пространството на латентните черти – едномерни (които предполагат, че манифестираните променливи реферират към една единствена латентна променлива) и многомерни модели (които боравят с пространства, формирани от няколко латентни променливи).

(2) Според вида на разпределението на латентната променлива в популацията – базирани на нормалното разпределение и на разпределения, които се отклоняват от него. Такива са например моделът на IRT, базиран на кривите на Рамзи (*Ramsay-curve item response theory, RC-IRT*) (виж Woods, 2006; 2007; 2008) както и модела, базиран на асиметрични разпределения (*Skew-Normal IRT*) *(*Bazan, Bolfarine & Branco, 2004).

(3) Според наличието на допускания относно вида на функцията, описваща връзката между латентната и манифестираната променлива – параметрични (PIRT) и (семейство) непараметрични (NIRT) модели.

(4) Според броя на параметрите на характеристичните криви на въпросите, включени в модела – едно-, дву-, три-, четири- и пет-параметрични модели.

(5) Според типа на функцията, свързваща изпълнението на тестовите въпроси от изпитаните лица и техните способности – модели, базирани на нормалната огива (*normal ogive models*) и логистични модели (*logistic ogive models*).

(6) Според формата на характеристичната крива, описваща връзката между изпълнението на тестовите въпроси от изпитаните лица и техните способности – модели със симетрични и с асиметрични криви. Така например Х. Базан и сътрудници предлагат семейство от IRT модели (*skew-normal IRT models*, *SN-IRT*), които базирани на асиметрично-нормално разпределение на способностите Θ и кумулативно асиметрично-нормално разпределение на характеристичната крива (Bazan, Bolfarine & Branco, 2004)

(7) Според типа на данните (според начина на скориране на отговорите) – бинарни (дихотомични) (правилен/ неправилен отговор), ординални (политомични, напр. при използване на скали от Ликертов тип) и континуални модели (Bolt, Cohen &

Wollack, 2001).

(8) Според равнището, на което се оценяват латентните черти – на индивидуално равнище (*Individual-level IRT*) или на групово равнище (*Group-level IRT*) (Bock & Mislevy, 1981; Mislevy, 1983, 1984; Tate, 1995).

В своя публикация Р. Хамбълтън прави подробен преглед на тестовите модели в рамките на IRT, които се използват в практиката на образователните и психологическите измервания (Hambleton, 1989).

Тук ще бъдат представени най-често използваните модели на IRT – едномерни, едно- до четирипараметрични, логистични, основани на нормалното разпределение на латентната променлива и на дихотомичните отговори на и. л.

Логистична функция

Логистичната функция е „четвърто поколение" функции, използвани за моделиране на връзката между скалата на латентните способности и манифестираните променливи, след линейната и стъпковата функция (от вида на тези, използвани при Гутмановите скали) и кумулативната нормална огива. Тя е предпочетена пред първите две поради по-високата си адекватност спрямо емпиричните данни, а пред третата – поради по-лесната си изчислимост. Заради тези нейни качества днес тя е сред най-често използваните вероятностни модели за представяне на тази връзка.

Логистичната функция определя семейството от логистични криви, обозначавани често като „криви на растежа", които се използват в различни области на знанието: в хуманитарните, социалните и стопанските науки (икономика), а преди това в природните (биологични) науки. Прилагат се за описание на дискретни, дихотомични (категориални) зависими променливи, които могат да приема две стойности: дадено събитие от пространство с две възможни събития се е осъществило (1) или не се е осъществило (0). В психологическите измервания пространството на възможните събития, което представлява интерес, е "изследваното лице е посочило ключовия отговор" и "изследваното лице не е посочило ключовия отговор".

Моделите, в които зависимата променлива е дихотомична, се разглеждат като нелинейни регресионни модели със зависима променлива, чието емпирично разпределение принадлежи към групата на експоненциалните разпределения. Особеност на тези модели е това, че вместо самата дихотомична зависима променлива, в уравнението се включва вероятността от поява на едно от двете събития (в образователните измервания - "изпитваното лице е отговорило правилно"). Моделираната връзка се представя като S-образна нелинейна регресионна крива от вида, представен на фигура 2 (Lord, 1980; Embretson & Reise, 2000).

Общият вид на логистичната функция е следният:

$$P(t) = \frac{1}{1+e^{-t}} \quad (52)$$

където $P(t)$ - зависима променлива

t - независима променлива

e - константа, основа на натуралния логаритъм, с приблизителна стойност 2.718

Еднопараметричният логистичен модел (1PL), както показва неговото наименование, включва само един параметър и това е трудността на въпросите b. Уравнението, изразяващо функционалната връзка между латентната променлива и вероятността от правилен отговор, има следния вид:

$$P_j(\Theta | b) = \frac{1}{1+e^{-D(\Theta - b_j)}};\ -\infty \le b \le +\infty \quad (53)$$

където:

$P_j(\Theta)$ - вероятност за правилен отговор на въпрос j

e = 2.718 - основа на натуралния логаритъм (неперово число)

Θ - личностов параметър (равнище на способности)

b_j - позиция/ трудност на същия въпрос

D - скалова константа, която при моделите с нормална огива има стойност 1.00, а при логистичните – 1.702

Скаловата константа D служи за конвертиране на логистичната крива в нормална огива, т.е. за преминаване от логистичен към нормален модел.

Моделът е разработен в началото на 60-те години от датския математик Г. Раш (Rasch, 1960), който, независимо от това, че тръгва от съвършено различни позиции, достига до логистичната форма на характеристичната крива. Поради това обикновено наименованията "еднопараметричен логистичен модел" и "модел на Раш" се използват като синоними.

Включването в модела само на параметъра на трудността b има дълбоки основания, тъй като това е единственият параметър, който е разположен на скалата на способностите, с която образува смесен континуум. Оттук теоретичните граници на изменение на параметъра b са същите, както и тези на Θ[5]: $-\infty \le b \le +\infty$, но практически рядко надхвърлят -3.00 $\le b \le$ +3.00.

Видно е, че шансът за успех на изпитваното лице се определя от дистанцията между неговото равнище на компетентност, представено като точка на континуума на способностите, и скаловата стойност на въпроса ($\Theta - b$). Останалите параметри на характеристичната крива присъстват в модела "невидимо" – дискриминативният параме-

[5] При въпроси с нулева дискриминативна сила стойността на b е неопределена.

тър е константна стойност (обикновено е фиксиран на равнище a = 1.00), параметърът за налучкване – на равнище c = 0.00, а параметърът за невнимателност – на равнище d = 1.00 за всички въпроси в теста.

Въпреки че еднопараметричният модел е най-опростеният сред останалите, той се радва на изключителна популярност, ако се съди по броя на теоретичните публикации или по невероятното разнообразие от сфери на практическо приложение. Този интерес се дължи на няколко негови характеристики, най-важната от които е наличието на достатъчен брой статистики, които позволяват оценката на параметрите на модела. По-конкретно, суровият тестов бал (изчислен като сумарна скала от дихотомичните отговори на въпросите) е достатъчна статистика за изчисляване на личностовия параметър Θ, а броят на правилните отговори на даден въпрос – достатъчна статистика за неговото локализиране (Harris, 1993). Той е единственият, при който лица с една и съща оценка на Θ_i имат и един и същ суров тестов бал. По този начин еднопараметричният модел, поне теоретично, се съгласува най-добре с бинарното скориране на въпросите и сближава измерването, основано на този модел, с измерванията в точните науки.

От друга страна, минимализирането на броя на параметрите и „лишаването" на логистичната крива, т.е. на въпросите, от други техни характеристики, отдалечава модела на Г. Раш от реалните практически ситуации и отслабва неговата обяснителна сила (Embretson & Hershberger, 1999; Embretson & Reise, 2000; Weiner et al., 2003). Много изследвания показват, че въпросите действително варират по отношение на дискриминативната си сила, което я прави реален и, вероятно, полезен параметър (Fan, 1998).

Двупараметричният логистичен модел (2PL) добавя нов параметър – дискриминативната сила на въпроса a, с което горното уравнение добива следния вид:

$$P_j(\Theta|\, a,b) = \frac{1}{1+e^{-Da_j(\Theta-b_j)}};\ -\infty \le a \le +\infty,\ -\infty \le b \le +\infty \qquad (54)$$

където:

a_j - наклон/ дискриминативна сила на въпрос j

Поради S-образната форма на характеристичната крива нейният наклон се променя с изменение на равнищата на компетентност. Той има максимална стойност в инфлексната ѝ точка, в която равнището на компетентност е равно на трудността на въпроса. Поради това дискриминативната сила на въпроса е свързана не толкова с общата форма на кривата, колкото с нейния наклон в точка $\Theta = b$. Действителната стойност на наклона в тази точка е a/4, но възприемането на a като наклон на характеристичната крива е приемлива апроксимация, която прави интерпретацията на този параметър по-лесна (Baker, 2001).

Добавянето на параметъра a води до това, че в общия случай всеки въпрос е представен от характеристична крива с различен наклон. Теоретичните граници на изменение на дискриминативния параметър са $-\infty \le a \le +\infty$, но практически той е ограничен между -2.80 $\le a \le$ +2.80 (Embretson & Hershberger, 1999, Weiner et al., 2003).

Трипараметричният логистичен модел (3PL) включва още един параметър – за налучкване на верния отговор c. Моделът е разработен от А. Бирнбаум, за да отрази едно обичайно обстоятелство в тестовата практика – изпитваното лице, което няма необходимата подготовка, прибягва до стратегия за налучкване на правилния отговор (Birnbaum, 1968). Поради това авторът разширява двупараметричния модел, като включва нов параметър, който да представи в логистичното уравнение влиянието, което отгатването оказва на вероятността от правилен отговор.

$$P_j(\Theta|\, a,b,c) = c_j + \left(1 - c_j\right)\frac{1}{1+e^{-Da_j(\Theta - b_j)}};\ 0 \le c \le +1.00 \qquad (55)$$

където:

c_j - долна асимптота/ вероятност от налучкване на правилния отговор

Третият параметър се дефинира чрез долната асимптота на логистичната крива, която вече не съвпада с абсцисната ос както при двупараметричния модел, а е издигната над нея ($P_j(\Theta) > 0.00$).

Тази промяна в конструкцията на характеристичната крива има две важни последствия. Първото е свързано с начина, по който се определя трудността на въпроса. Той вече не е локализиран на позиция, съответстваща на вероятност $P(\Theta)$ = 0.50. Тъй като долната граница на характеристичната крива е c, трудността се дефинира като точка на скалата Θ, съответстваща на вероятност от правилен отговор:

$$P\left(\Theta\right) = c + \frac{1}{2}\left(1 - c\right) = \frac{1+c}{2} \qquad (56)$$

Това е средата на интервала между c и 1.00, което при c = 0.00 е точно 0.50.

Втората последица е концептуална – моделът на А. Бирнбаум вече не е логистичен, макар и по традиция да се приема за част от семейството на логистичните модели (Baker, 2001).

Важно е да се отбележи, че, за разлика от b, стойността на c не се променя с изменение на равнищата на компетентност, т.е. индивидите, позиционирани като точки дори и на противоположните краища на континуума Θ, имат еднакъв шанс да посочат верния отговор, прибягвайки до отгатване. Параметърът c може да варира в интервала 0.00 $\le c \le$ 1.00, но практически стойности над 0.10 се считат за неприемливи. Дискри-

минативната сила се интерпретира по същия начин, както и при предходния модел – като пропорционална на наклона на логистичната крива в точка $\Theta = b$, но тук действителната стойност на наклона е $a(1 - c)/4$.

Авторът на последните два модела А. Бирнбаум (Birnbaum, 1968) посочва една от техните слабости – липсата на достатъчно статистики, за да се направят оценки на параметрите на въпросите, независима от личностовия параметър Θ (Lord, 1980; Embretson & Hershberger, 1999; Baker, 2001; Weiner et al., 2003; Downing & Haladyna, 2006).

Четирипараметричният логистичен модел (4PL) включва, освен трите параметъра на въпросите, описани по-горе, и параметъра d - "невнимателност". Моделът е разработен от М. Бартън и Ф. Лорд, за да отчетат влиянието на още един фактор върху изпълнението на тестовите въпроси, оползотворявайки по-пълно потенциала на логистичната крива (Barton & Lord, 1981). Параметърът се изчислява по следната формула:

$$P_j(\Theta|\, a,b,c,d) = c_j + \left(d_j - c_j\right)\frac{1}{1+e^{-Da_j(\Theta-b_j)}};\ 0.00 \le c < d \le +1.00 \qquad (57)$$

където:

d_j - горна асимптота/ невнимателност

Този параметър, както бе отбелязано, отразява по-слабата мотивация у изпитваните, тяхното нежелание за постигане на високи постижения, нехайство или пък допускането на грешки поради небрежност при маркиране на отговорите. Четвъртият параметър се дефинира чрез горната асимптота на логистичната крива, която тук не съвпада с горната абсциса, а е разположена по-ниско от нея. Така вероятността от правилен отговор на индивидите, по-конкретно на тези с високи нива на способности, става $P(\Theta) < 1.00$ при нарастване на стойността на d (Embretson & Reise, 2000).

1.2.5. Предимства и недостатъци на IRT

“Тихата революция”, извършена в областта на психологическите измервания чрез създаването и развитието на IRT, има своето логично обяснения – това е теоретична рамка, която осигурява по-добре от останалите теории постигането на основната цел на измерването – получаване на такива оценки на личностовия параметър (позицията на и. л. на континуума Θ), които да бъдат неизместени, състоятелни и надеждни (Гласс и Стэнли, 1976). Цялата верига от теоретични конструкти в IRT и методите за тяхната оценка водят към тази цел. Всички изследователи, работещи в областта на IRT, са единодушни по отношение на нейните безспорни качества, които, в сравнение с нейния предшественик – Класическата теория, са и безспорни предимства (Hulin,

Drasgow & Parsons, 1983; Hambleton, Swaminathan, & Rogers, 1991; Embretson & Reise, 2000; Harvey, 2003).

Новата теория предлага по-задълбочено разбиране на каузалната връзката между индивидите и айтемите, между латентните и манифестираните променливи. За разлика от Класическата теория, която отказва да се занимава с този въпрос, в IRT се предлага вероятностно моделиране на тази връзка, като се разглежда вероятността от коректен отговор на всеки тестов въпрос. Този подход дава възможност за оценка на характера на тази връзка въз основа на извадки, които не покриват целия диапазон на континуума на способностите, а само част от него.

Новата теория предлага детайлна информация за функционирането на отделните въпроси, разгръщайки широка система от техни характеристики. Независимо от наличието на някои общи концепции за въпросите, които тя споделя с Класическата теория, IRT предлага изцяло нова математическа интерпретация на тези характеристики, разглеждайки ги като параметри на функционалната връзка между латентните и манифестираните променливи. Както беше отбелязано по-горе, дори и непредставителни извадки могат да послужат за формиране на неизместени оценки на тези параметри.

IRT въвежда и други нови конструкти като характеристичната крива на теста, която отразява функционалната връзка между скалата на способностите Θ и действителния бал на индивидите. Тя играе важна роля през етапа на интерпретиране и представяне на резултатите от теста, като служи за преобразуване на скалата на способностите в скала, която е свързана с тестовия бал в Класическата теория и поради това е лесно разбираема за потребителите на тестовите резултати.

Специфични за IRT са и концепциите за информационните функции на въпросите и на теста. Първият конструкт отразява прецизността на измерването в термините на отклонението на оценките на личностовия параметър от действителната им стойност. Подобно на други конструкти в IRT, информативността (прецизността) на измерването се разглежда не като обща оценка за всички нива на способност, а като променлива, която може да бъде оценена за всеки тестов въпрос, за всяка точка на континуума Θ_i. Информационната функция на теста представлява зависимостта между скалата на компетентността Θ и количеството информация, определена за всяка нейна точка въз основа на всички въпроси в теста.

Като мярка за определяне на прецизността, т. е. на отклоненията на оценките $\hat{\Theta}_i$ от стойността на търсения личностов параметър Θ_i, се използва стандартната грешка на оценката (*standard error of estimation*). За оценка на действителните стойности на Θ в IRT най-често се използва методът на максималното правдоподобие (*maximum likelihood*), който представлява итеративна процедура за повишаване на прецизността на оценяването на личностовия параметър, основана на съпоставяне на

актуалните стойности на отговорите на индивидите на въпросите от теста и асоциираните с тях вероятности за правилен отговор.

Не на последно място, IRT дава възможност за оценка на консистентността на отговорите на групи от изпитани, формирани по даден признак (социо-демографски или друг) на равнище айтем и цялостен тест чрез наблюдение на диференциалното функциониране на отделните айтеми (*differential item functioning, DIF*) или диференциалното функциониране на теста (*differential test functioning, DTF*). Това специфично „поведение" на айтемите или теста, при което тези единици се отличават с различни характеристики при различните групи от индивиди при контролиране на общото равнище на способностите в тези групи. От гледна точка на новата теория, DIF се проявява в случаите, в които лица от различни групи, характеризиращи се с едно и също равнище на способности Θ, отговарят позитивно на даден въпрос с различна вероятност. Следователно тази особеност на айтемите може да се прояви чрез различните стойности на параметрите на въпросите, оценени чрез IRT, при различните групи от индивиди.

Едно от най-стойностните предимства на новата теория е това, че тя прекъсва две важни зависимости, които, от своя страна, са „вродени" недостатъци на Класическата теория. Първата от тях е зависимостта на статистиките на въпросите от групата от и. л., въз основа на която са получени. Втората е зависимостта на тестовия бал от конкретния набор от въпроси (тест). Параметрите на въпросите и личностовия параметър в рамките на IRT са независими от средата и в този смисъл са инвариантни (Lord, 1980; Hambleton & Swaminathan,1984; Hambleton, Swaminathan, & Rogers, 1991; Fan, 1998; Baker, 2001). Тази особеност лежи в основата на решаването на редица важни практически задачи като прилагането на процедурата на изравняване/ „свързване" на различни тестове, както на развитието на един модерен подход за оценяване, известен като Компютъризирано адаптивно тестиране (*Computerized adaptive testing, CAT*) (Hambleton, Swaminathan, & Rogers, 1991).

Подобно на класическата, и новата психометрична теория не е лишена от недостатъци. След серията от положителни оценки за нейните качества, не е учудващо, че критиките към нея като към теоретична конструкция са много малко.

Ф. Бейкър отбелязва, че с включването в трипараметричния модел на параметъра за налучкване на коректния отговор, логистичната функция „е загубила някои от добрите си математически качества" (Baker, (2001, стр. 28) и от тази гледна точка трипараметричният модел не е логистичен, макар и да се третира като такъв. В допълнение, параметърът за налучкване (c), както и този за невнимателност (d) не се изменят функционално по протежение на континуума Θ, поради което и. л. с различни нива на способности се характеризират с равен шанс да посочат (или да не посочат) коректния отговор.

Най-съществените критики към IRT обаче са насочени към валидността на нейните допускания (по-скоро – тези на отделните теоретични модели). Много изследова-

тели, между които и Д. Харис, посочват като основна слабост на новата теория това, че трите логистични (1, 2 и 3-параметрични) модела "…допускат, че една единствена едномерна черта лежи в основата на данните - допускане, което рядко, ако изобщо някога, се среща в реалните тестови ситуации" (Harris, 1993, стр. 161). Освен допускането за едномерност, обект на критики е и допускането за нормално разпределение на латентната/ латентните променливи. Анализът на голям масив от тестови данни от различни източници дава основание на Т. Мичери да заключи, че твърде малка част от разпределенията са „дори сравнително близка апроксимация на Гаусовото" (Micceri, 1989, стр. 161). На мнение, че нормалното разпределение на тестовите балове е много рядко явление, са и други изследователи (Nunnally, 1978; Brown, 1996). По един особено драматичен начин тези мнения, формирани в повечето случаи от теоретична гледна точка, са обобщени от Р. Гиъри по следния начин: „Нормалността е мит; никога не е имало и никога няма да има нормално разпределение." (Geary, 1947, стр. 241).

К. Фан отбелязва, че тезата за инвариантност на статистиките на айтемите, макар и да е в сърцевината на новата теория, не бива да се надценява, защото, ако не бъде доказана, сложността на моделите на IRT ще бъде лишена и от теоретично, и от практическо основание (Fan, 1998).

Като друг недостатък може да се посочи допускането, че в рамките на даден модел, характеристичната крива има една и съща форма за всички айтеми. Това допускане е също твърде силно, ако се вземат предвид резултатите от методите за оценка на годността на съответния модел, например за броя на параметрите, които описват характеристичната крива на въпросите. Прилагането тези методи почти винаги води до отхвърлянето на по-малко или повече айтеми, които не съответстват на избрания модел, независимо от това дали той е 1-, 2- или 3-параметричен. Това означава, че по правило айтемите в един и същи тест могат да бъдат описани с характеристични криви в рамките на различни модели.

От прагматична гледна точка може да се посочи и това, че за оценка на допусканията на избрания модел, както и за цялостна оценка на неговата годност, са необходими по-големи извадки (Fan, 1998). По-големи извадки са необходими и за постигане на състоятелни и надеждни оценки на неизвестните параметри в етапа на пилотното тестиране, особено тези на тестовите въпроси. Тази особеност също предполага по-големи усилия или по-ограничение възможности за прилагането на моделите на IRT. Резюмирайки предимствата и недостатъците на новата психометрична теория, бихме могли да отбележим нейните безспорни достойнства като теоретична система, които, обаче, биха се проявили трудно при съприкосновение с реалните ситуации.

2. Преглед на изследванията по проблема за приложимостта на тестовите теории и техните модели

2.1. Общи наблюдения

Целта на този обзор е да се представи състоянието на изследванията върху приложимостта на Класическата тестова теория (CTT) и Теорията за отговор на тестов въпрос (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 пестри 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 психометрич-

ното 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 много-

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

2.2. 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 извли-

ча 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 про-

цедурите 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 въпроса), 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[6]. 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 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 из-

[6] 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.

ползват 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 в интервала 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 нега-

тивни, 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 ASVAB (*Armed services vocational aptitude batery*) 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 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 $r_{\text{verbal}} = 0.39$, $r_{\text{nonverbal}} = 0.69$ и $r_{\text{total}} = 0.71$. 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 CTT, 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 CTT 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 CTT. 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 CTT (Nukhet, 2002), а според К. Фан такава съг-

ласуваност 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 CTT 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 CTT и IRT при анализа на въпросите от теоретичната част на теста за придобиване на свидетелство за управление на МПС 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 CTT. 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 CTT 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_{pbis} (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. Аседоин 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 CTT, 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.

2.3. 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 комуникаци-

онните 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. Т. Мичери 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, но 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 Га-

усовото 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 графичния тест 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 (виж уравнение 55) 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 дименсии (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". 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, базиран на Кумбсовата теория на данните (*Unfolding IRT*) за измерване на културни норми (Coombs, 1964; 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.

II. Обща постановка на изследването

1. Обосновка на необходимостта от изследване на приложимостта на тестовите теории

Представяното изследване е проведено в контекста на засиления интерес към теоретичните и приложните аспекти на психологическите измервания в поведенческите, социалните и хуманитарните науки и, през последните години – и в медицинските науки. То е фокусирано върху две психометрични теории, които, без съмнение, са доминиращи в тази научна област: Класическата тестова теория (CTT) и Теорията за отговор на тестов въпрос (IRT). 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.

Една от основните сфери на приложение на двете психометрични теории са тестовете за постижения, 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. Тя обаче, както беше показано в предходната глава на работата, се отличава с редица несъвършенства. Това обстоятелство повдига важния въпрос за необходимостта от замяна на стария психометричен модел с новата Теория за отговор на тестов въпрос при конструирането и обработването на резултатите от ТОП. Използването на IRT, поради нейните безспорни теоретични предимства пред CTT, би довело до съществено подобряване на измерителните качества на теста.

Преди да представим аргументите в полза на необходимостта от провеждане на такива изследвания, е необходимо да разгледаме 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 латен-

тни променливи, съответно наблюдаван и действителен бал, трудност на въпроса или неговата характеристична крива, както и отношенията (връзките) между тези понятия. 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, най-често 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 тяхното приложение. Поради това един теоретичен модел следва да се проверява от гледна точка на неговото съответствие с реалните феномени, към които предстои да 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 съответст-

ват 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, които 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 об-

ръща 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 способност.

Същевременно CTT и IRT са две твърде различни системи за скалиране. Те са 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 конструктите, свързани с описанието на психометричния инструмент на ниво тестов въпрос, субтест или цялостен тест, така и конструктите, отразяващи измерваната способност.

Както беше отбелязано, в специализираната литература се наблюдава дефицит 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 прилични снимки, но с малък разход на енергия. IRT е съвършено нов модел апарат, който прави ярки, живи и контрастни снимки, но с голям раз-

ход 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).

2. Цели

Основната изследователска цел в представената разработка е да се направи сравнително изследване на приложимостта на двете основни психометрични теории – Класическата тестова теория и Теорията за отговор на тестов въпрос, върху реални данни от Теста за общообразователна подготовка. Концепцията за приложимостта ще бъде разгледана в два нейни аспекта:

(а) степента на съответствие между допусканията на теоретичния модел и характеристиките на емпиричните тестови данни, т.е. валидността на основните допускания на теоретичния модел спрямо данните;

(б) степента, в която очакваните свойства на теоретичния модел се проявяват в емпиричните тестови данни, най-вече по отношение на очакваното „поведение" на статистиките на тестовите въпроси.

В светлината на разгледаните предимства и недостатъци на двете тестови теории, в изследването ще бъдат потърсени отговорите на следните два основни изследователски въпроса:

(1) Съответства ли Класическата тестова теория, в рамките на която функционира ТОП, в достатъчно висока степен на тестовите данни?

(2) Би ли довела замяната на стария психометричен модел с новата Теория за отговор на тестов въпрос, поради нейните безспорни теоретични предимства, до подобряване на измерителните качества на теста?

Тук следва да се подчертае, че тежестта на изследването е върху търсенето на различни по характер свидетелства „за" и „против" приложението на новата психометрична теория при разработването и анализа на резултатите от ТОП.

Изследването е фокусирано върху проверка на адекватността на следните два теоретични модела:

В теоретичната рамка на Класическата тестова теория - едномерен, с нормално разпределение на действителния бал, τ-конгенеричен модел. Както е известно, обикновено СТТ се разглежда като единна теория, която не включва в себе си различни модели. Някои автори (виж Steyer, 2001; DeVellis, 2003) дефинират три нейни модела в зависимост от отношенията между които и да е два теста X_i и X_j, които се използват за оценка на една и съща индивидуална характеристика и в този смисъл са взаимозаменяеми. τ-конгенеричните тестове се дефинират от допускане (12) за τ-конгенеричност (за разлика между действителните балове на двата теста, които корелират помежду си линейно с коефициент $b \neq 1$) и (13) за липса на корелация между грешките на измерването в двата теста. В теоретичната рамка на Теорията за отговор на тестов въпрос - основан на дихотомични отговори, едномерен, с нормално разпределение на латентната способност Θ, параметричен, логистичен модел.

За по-кратко тук и по-нататък в текста ще реферираме към тези модели като към „базови модели" със значение на възможни, първоначални рамки за описание на данните от ТОП.

Изборът на тези два базови модел е обоснован от множество теоретични предпоставки, представени в предходната глава на тази разработка. Тук ще добавим още няколко аргумента от прагматичен характер. Дефинираният τ-конгенеричен модел се базира на най-слабите допускания на „меката" Класическа тестова теория, по-специално на най-„слабия" тип корелационна връзка между действителните балове на два паралелни теста. Диференциалните признаци, описващи базовия модел на IRT, също са сред най-често срещаните както при нейното теоретично представяне в литературните източници, така и в нейното прилагане в различни емпирични контексти. Следователно може да се очаква, че тези модели ще бъдат в най-висока степен адекватни и към данните от ТОП.

По-съществения аргумент в полза на избора на тези два модела обаче е самият Тест по общообразователна подготовка. Неговото предназначение, технология на конструиране и скалова структура, начинът обработване на суровите данни, на интерпретация и използване на получените резултати водят към дефинирането на тези първоначални модели.

В светлината на тези предпоставки в изследването ще бъдат поставени и след-

ните две групи от изследователски въпроси, които конкретизират формулираните по-горе общи проблеми.

Първата група произтича от разбирането на концепцията за приложимостта като съответствие между допусканията на теоретичния модел и характеристиките на тестови данни:

(1) Каква е формата на разпределенията на латентните способности? Доколко обосновано е допускането, че латентните способности следват нормалното Гаусово разпределение?

(2) Каква е размерността на пространството на латентните способности, които обуславят отговорите на и. л. на въпросите от ТОП? Има ли емпирични свидетелства, които да подкрепят допускането за тяхната едномерност?

Втората група е свързана с разбирането на приложимостта като проява на очакваните свойства на теоретичния модел в емпиричните тестови данни:

(3) При кой от двата теоретични модела, съответно на CTT и IRT, статистиките на тестовите въпроси са инвариантни в различни условия, т. е. при различни извадки от индивиди?

(4) Доколко статистиките на тестовите въпроси, определени в рамките на един и същи модел, в едно и също условие, т. е. при една и съща извадка от индивиди, функционират независимо една от друга?

(5) Наблюдава ли се съгласуваност между индексите, определени в рамките на CTT, и съответните им параметри, определена в рамките на IRT, в едно и също условие, т. е. при една и съща извадка от индивиди.

Ще бъде потърсен отговорът и на още един въпрос, свързан с базовия модел на IRT:

(6) Дали дефинираният едномерен, с нормално разпределение на латентната способност Θ базов модел на IRT съответства на данните от теста и ако не съответства, кой модел би бил по-подходящ за приложение?

В процесуален план замисълът на изследването може да бъде обобщен в следния вид: (1) да се определят два първоначални модела в рамките на CTT и IRT; (2) да се проведат специфични анализи (в частност, статистически тестове) за изследване на различни аспекти на приложимостта на базовите тестови модели към реалните данни; (3) да се съберат различни видове доказателства за степента, в която данните удовлетворяват някои от основни допускания на моделите; (4) да се изследват различни видове взаимовръзки между едноименните статистики (например индекс на трудност p в рамките на CTT и параметър на трудност b в рамките на IRT) и разноименните статистики на въпросите (например индекси на трудност p и на дискриминативна сила D в рамките на CTT) от гледна точка на особеностите на тези статистики, които произтичат от съответния модел; (5) въз основа на събраната информация да се изгради адекватен модел (или модели) на IRT.

3. Методология

3.1. Основни изследователски подходи

Формулираната основна цел, както и съпътстващите я изследователски въпроси, са свързани с различни аспекти на приложимостта на психометричните теории към конкретна база от емпирични данни, които, следва да отбележим, не покриват тази проблематика в цялата и широта. Съчетанието от разнородни проблеми предполага провеждането на множество относително самостоятелни изследвания. Те са предназначени за постигането на различни конкретни изследователски цели, но това, което ги обединява е, че всяко изследване ще има своя дял в изясняването на комплексната проблематика, свързана с приложимостта на психометричните модели.

Друг аспект, който споява отделните изследвания в рамките на настоящата разработка, е прилагането на три основни изследователски подхода.

Първият от тях е познат като Изследователски анализ на данни (*Exploratory data analysis, EDA*), предложен и обоснован от Дж. Тюки (Tukey, 1977; Hartwig & Dearing, 1979; Behrens, 1997). Появата на този подход е логична последица от наблюденията, че често пъти при изследване на даден феномен и при наличието на големи съвкупности от данни, изследователят не разполага с достатъчно информация, за да формулира смислена хипотеза или да изгради модел, който след това да подложи на проверка. Схващането на Дж. Тюки е, че в тези ситуации по-полезното действие е последователното и многостранно изследване на събраните данни, което да доведе до „извличане" на модела от тях, вместо да бъде предпоставен.

„Изследователски анализ на данни е подход за учене от данните, предназначен за разбиране на света"- отбелязват Ъ. Уайнър и сътрудници (Weiner et al., 2003, стр. 34). За разлика „класическия" потвърдителен подход (*Confirmatory data analysis*, *CDA*), основан на проверка на предварително формулирани хипотези (или модели) чрез съответните статистически тестове, изследователският анализ предполага системно изследване на взаимовръзките между (големи) групи от променливи. Дж. Тюки застава на по-широки методологически позиции, приемайки, че боравенето с данните включва много повече по-разнообразни методи и средства, отколкото традиционното формулиране и проверка на хипотези, на които се отдава прекалено голямо значение. Целите на изследователя, използващ EDA, са „…да разкрие неочакваното, да не се остави да бъде излъган [от данните] и да разгърне съдържателни описания" (Weiner et al., 2003, стр. 36). Дж. Тюки сравнява работата на изследователя с детективската работа, свързана с детайлно изследване на обстоятелствата, разкриване на връзки (модели) и генериране на първоначални хипотези, след което следва верифициране на хипотезите и потвърждаване на моделите.

EDA е методологически подход (дори изследователска "нагласа"), а не набор от

конкретни (статистически) техники. В неговите рамки могат да бъдат решавани разнообразни задачи като разкриване на устойчиви модели, редуциране на данните, проверка на допускания, разкриване на латентни структури и др. В по-общ план логическата последователност от действия при прилагане на този подход е следната:

формулиране на изследователски проблем ⇨ събиране на сурови данни ⇨ анализ на данните ⇨ изграждане на модел ⇨ извеждане заключения за данните

Вторият подход е свързан с осигуряването на вътрешната валидност на резултатите от изследванията. Както беше отбелязано, в рамките на EDA за изследване на даден феномен могат да се приложат различни изследователски методи с различна степен на адекватност. Поради това използваната методологията следва да бъде резултат от обоснован избор от на изследователя. Понятието „вътрешна валидност" се свързва обикновено с изследванията на причинно-следствените отношения между променливите и обозначава истинността на предположението (извода), че наблюдаваният ефект в зависимата променлива действително е породен от независимата променлива. Някои автори му придават по-тясно значение, включвайки в него адекватността на използваната методология. Други автори предпочитат термина „статистическа валидност" или „валидност на статистическите изводи" (*statistical conclusion validity*) (Cohen & Swerdlik, 2005). Т. Кук и Д. Кемпбел дефинират понятието още по-тясно, свеждайки го до коректността на изводите за наличие на ковариация между независимите и зависимите променливи при дадено равнище на α и на наблюдаваните дисперсии (Cook & Campbell, 1979). В настоящата разработка ние ще се придържаме към по-широкото тълкуване на Р. Коен и М. Суердлик, според които статистическа валидност се отнася до адекватния подбор на статистическите методи и до коректността на последващата интерпретация на получените чрез тях резултати (Cohen & Swerdlik, 2005).

Третият подход е свързан с осигуряване на външната валидност на резултатите от изследванията. Тя се отнася до това дали и в каква степен резултатите от едно изследване и направените от тях изводи могат да бъдат генерализирани върху други ситуации, т.е. доколко могат да се пренесат и върху други, нереализирани изследвания от същия вид. От съществено значение е до каква степен направеното изследване е представително за съвкупността от всички възможни изследвания, най-вече от гледна точна на представителността на извадките.

В настоящото изследване можем да говорим за два типа извадки: (а) на изпитаните лица, които представляват част от генерална съвкупност, която може да бъде определена като периодна (включваща кандидат-студентите, които се явяват на ТОП през кандидат-студентските кампании в НБУ) или по-скоро като динамична (променяща състава си през времето, включваща всички потенциални кандидат-студенти, завършили средно образование) и (б) на тестовите въпроси, включени в съответния вариант на теста. Извадките от индивиди могат да бъдат разгледани като представителни (поради естествено случайния начин на разпределение на кандидатите между от-

делните изпитни сесии и тестови варианти, както и от сравнително големият им обем). Въпреки това тези извадки не са случайни в статистически смисъл и се различават значително по своите обеми. Що се отнася до извадките от въпроси, те не са случайни (поради процедурите на селекция на въпросите във фазата на конструиране на тестовите варианти) и са сравнително малки по обем, особено на субтестово равнище.

Поради това като мярка за осигуряване на външната валидност е приложен подходът за репликиране на ситуациите, т. е. за паралелно изследване на множество от еднотипни съвкупности от данни чрез извършване на съответните анализи върху множество от тестови варианти, използвани в различни точки от време. А това означава наблюдение на различни извадки от индивиди и от тестови въпроси.

Тясно свързан с проблема за външната валидност е и проблемът за осигуряване на екологичната валидност на резултатите. Понятието за екологична валидност е въведено от Е. Брънсуик във връзка с изследванията върху възприятията, които се провеждат в лабораторни условия (Brunswik, 1956, по Зиновиева, 2009). Авторът обръща внимание на това, че е необходимо разработването на такива модели, които описват възприятията не в повече или по-малко изкуствена среда, а в реални житейски ситуации. Идеята е подета от Дж. Гибсън, който развива становището, че резултатите от един лабораторен експеримент могат да се различават съществено различни от тези, получени в реална среда до степен да се окажат невалидни по отношение на явленията, които обясняват (Gibson, 1979, по Зиновиева, 2009). Екологичната валидност, следователно, е степента, в която един модел, метод или резултатите от дадено изследване могат да бъдат отнесени към реалните ситуации, в които функционира индивидът (Зиновиева, 2009).

При изследване на въпросите, свързани с приложимостта на теоретичните психометрични модели, алтернативата е, както нееднократно бе подчертано в предходните глави на разработката, използването на симулирани данни срещу използването на реални данни. В представеното изследване екологичната валидност на резултатите е обезпечена чрез използването на реални данни, получени при множество реални изпитни ситуации.

3.2. Дизайн

По-голяма част от изследванията ще бъдат направени в парадигмата на корелационните анализи. Този изследователски подход, както е известно, не е свързан с манипулиране на някои от променливите и с търсене на ефекта от манипулацията върху други променливи, а със системни наблюдения върху взаимовръзките между групи от променливи. Поради това тези изследвания включват само „зависими" променливи – тези, които са наблюдавани и регистрирани. Доколкото обаче дихотомията независими – зависими променливи се използва терминологично главно в експерименталните изследвания, тук и по-нататък в текста ще използваме контекстно-

обусловена терминология: действителният тестов бал τ в концептуалната рамка на CTT, както и способностите Θ в IRT, са обяснителни, ненаблюдаеми или латентни променливи, които играят роля на независими променливи (предиктори). Векторите, формирани от отговорите на отделните въпроси, суровият тестов бал от отделните субтестове, както и общият тестов бал, са наблюдавани или обяснени променливи, които играят ролята на зависими променливи. Като променливи в някои от изследванията ще бъдат включени статистиките на въпросите: индексите на трудност p и на дискриминативна сила D в теоретичната рамка на CTT, както и параметрите дискриминативна сила a, на трудност b и на налучкване на коректния отговор c в рамките на IRT.

3.3. Източник на данни

Основен източник на данни в емпиричното изследване е Тестът по общообразователна подготовка (ТОП), който стои в основата на приемните процедури в Нов български университет. Прототипът на ТОП е апробиран пилотно за пръв път през 1995 г. като приемен изпит за бакалавърската програма “Икономика” в тогавашния Свободен факултет. ТОП е замислен като тестова батерия с осем относително самостоятелни части (раздели): Български език, Литература, История, География, Математика, Физика, Химия и Биология. Те са базирани върху учебното съдържание на основните хуманитарни и природо-математически дисциплини, които се изучават в българското средно училище и формират общообразователната подготовка на зрелостниците. Всеки раздел съдържа по десет тестови въпроса с множествен избор, чието съдържание не излиза извън рамките на учебната програма за задължителна и задължителноизбираема подготовка. Въпросите в теста имат по-скоро приложна, отколкото теоретична насоченост. Като цяло те не предполагат дословното познаване на дефиниции, формули или правила от учебниците, но владеенето им би могло да бъде предпоставка за по-добро ориентиране в проблемите и за намиране на правилните решения. Част от въпросите са свързани и с решаване на практически проблеми, разчитане на графики и таблици.

През 1996 г. бяха проведени редовни кандидатстудентски изпити за над 10 бакалавърски програми на Свободния факултет. За тази цел бяха разработени 8 варианта на ТОП с различен обем, структура и съдържание, съобразени със специфичните изисквания на програмите. В съответствие с решенията на общоуниверситетската конференция на тема "Приемът на студенти в НБУ - проблеми и процедури", проведена през есента на 1996 г., в концепцията за структурата и прилагането на теста като приемен изпит бяха направени някои промени, насочени към уеднаквяване на приемните процедури за всички учебни програми. Важно решение на тази конференция бе утвърждаването на ТОП като единен, стандартен общоуниверситетски тест за прием на студенти, независимо от програмата, в която кандидатстват.

В изпълнение на препоръките на конференцията, за кандидатстудентските из-

питни сесии през 1998 г. в ТОП бяха включени още два раздела - Разсъждения и Семантика с по 10 тестови въпроса. Чрез тях се подлагат на проверка умения и способности, които не са обвързани с конкретен учебен предмет, но са особено важни за успешното обучение на студентите и се придобиват не само в училище, но произтичат и от житейския опит на човека. Въпросите в раздел Разсъждения са базирани на адаптирани затворени текстове от различни източници и с разнообразна тематика. Чрез тях се проверяват уменията на кандидат-студентите за разбиране при четене, за организиране и анализиране на информацията в тях и за извеждане на обосновани изводи от нея; способността за аргументиране на предположения; уменията за логическо и критично мислене. Въпросите в раздел Семантика са предназначени за проверка на езиковата компетентност на кандидат-студентите на по-ниско ниво - морфема, дума, фраза, устойчиво словосъчетание. Съдържанието им обхваща значението на езиковите единици, смисловите отношения, синонимия и антонимия, класификация, словообразуване и др.

Ако разгледаме администрирането на ТОП в хронологичен план, ще се убедим в големите мащаби на неговото приложение. За 15-годишен период на провеждане на кандидат-студентски кампании в НБУ (до 2010 г) изпити са били проведени на 100 дати, разработени са 314 варианта на теста и са били изпитани 156 213 кандидат-студенти. По-подробна информация по години е представена в Приложение 1.

Профилът на ТОП може да бъде дефиниран като широкоспектърна, неспециализирана тестова батерия за подбор на студенти въз основа на тяхната общообразователна подготовка от средното училище. Тази обща характеристика следва да бъде конкретизирана чрез определяне на типа на теста по три диференциални признака, представени в следващата таблица:

Таблица 1. Диференциални признаци за категоризиране на ТОП

цели	удостоверяване	подбор
компетентности	постижения	склонности
статус	норма	критерий

(1) Според целите на провеждане на изпита ТОП е тест за подбор на лица за продължаване на обучението им в следваща образователна степен.

(2) Според способностите, които са обект на оценяване, ТОП принадлежи към групата на тестовете, предназначени за измерване на постижения (*achievement tests*). Това са най-широко разпространените тестове, предназначени за измерване на знания и умения в определени предметни области, усвоени при известни и контролирани условия. Типологично те се противопоставят на тестовете за измерване на способности/ склонността за обучение (*aptitude tests*), базирани на общия познавателен опит на ин-

дивида и предназначени за предсказване на неговите постижения при бъдещо (предстоящо) обучение. При приложението на двата типа тестове обаче не може да се очертае някаква рязка граница. Наблюдава се определено смесване на техните функции, включително и използване на тестовете за постижения като предиктори за степента, в която индивидът би се справил успешно в новата образователна среда. Поради това двата термина все по-често се заменят с „..по-неутралния термин способност (*ability*) в названията на средствата за оценка на когнитивното поведение" (Анастази и Урбина, 2001, стр. 516-517). По този диференциален признак ТОП е тест за постижения, натоварен с функции за измерване на склонността към обучение и за предсказване на бъдещи постижения, т.е. тест за способности.

(3) Според начина на определяне на статуса на изпитания ТОП принадлежи към категорията на тестовете, ориентирани към норма (*norm-referenced tests*), които са предназначени са за определяне на относителния статус на един индивид спрямо други индивиди по отношение на измервания признак. Този тип тестове се основават на съпоставяне на индивидуалния резултат на изпитания с резултатите на представителна извадка от индивиди, които са на същото образователно равнище (група „норма") или с резултатите на изпитаните от същата група, изпълнила теста.

Въз основа на тези диференциални признаци ТОП може да бъде определен като тестова батерия за подбор на лица, които да продължат обучението си в следваща образователна степен, основан на техните постижения по дисциплини, изучавани в предходната степен, чрез определяне на относителния статус на изпитаните. В прагматичен аспект всички тези особености се използват в приемните процедури на НБУ, като за балообразуване се използват както резултатите от отделните раздели (субскали) на теста, така и общия тестов бал.

За осъществяване на целите на изследването бяха подбрани 15 варианта на ТОП от различни кандидатстудентски кампании през 6-годишен период, от различни етапи на тези кампании, като броят на и. л. варира в сравнително широки граници от 454, явили се на вариант 146, до 1019, явили се на вариант 192. По-подробни данни са дадени в Приложение 2.

3.4. Изследвания и процедури

За постигане на основната цел на изследването, както и на произтичащите от нея изследователски въпроси, ще бъдат проведени серия от относително самостоятелни емпирични изследвания. В частност ще бъдат проучени следните пет аспекта на поставения общ проблем.

Първите две изследвания са свързани с разбирането на концепцията за приложимостта като съответствие между допусканията на теоретичния модел и характеристиките на тестови данни. На оценка ще бъде подложена валидността на две основни допускания: за едномерност на латентните структури и за нормалност на разпределе-

нието на латентните променливи (изследователски въпроси 1 и 2).

Изследване 1. Анализ на формата на разпределенията на латентните променливи

Изследването ще бъде направено главно с оглед на верифициране на допускането за тяхната нормалност. То е дефинитивно за по-голяма част от моделите на IRT, включително и за базовия модел. По-голяма част от тестовите статистики в рамки на CTT са статистики на нормалното разпределение или са базирани на него. В допълнение, всички методи за оценка на надеждността и валидността на резултатите, базиращи се на линейни (регресионни) модели, също предполагат нормалност на разпределението на латентната променлива (Graham, 2006). Анализът на формата на разпределенията ще бъде направена върху данните от 12 тестови варианти, на две нива – субтестово и тестово. Ще бъдат приложени различни числови и графични техники - статистически тестове на съгласието, определяне на индексите на асиметрия и ексцес на съответните разпределения, хистограми и нормални вероятностни графика.

Изследване 2. Анализ на латентните структури на Теста по общообразователна подготовка

Изследването ще бъде направено главно с оглед на верифициране на допускането за тяхната едномерност. Това допускане е експлицитно за по-голяма част от моделите на IRT, но дори и при прилагането на многомерни модели е необходимо първоначално да бъде установена ясно размерността на латентното пространство (Embretson & Reise, 2000). Макар че при CTT такова изискване обикновено не се дефинира експлицитно, някои автори разглеждат класическия модел на измерване като „предназначен за измерване на един-единствен феномен" (DeVellis, 2003, стр. 28). Някои от основните конструкти в CTT, както и методите за тяхната оценка, също се базират на допускането за едномерност на латентния признак.

Проверката на допускането за едномерност на латентните структури на ТОП ще бъде направена върху данните от 12 тестови варианти, на две нива – субтестово и тестово. Като основен метод за установяване на размерността ще бъде използван изследователския анализ на главни фактори. За определяне на релевантния брой на факторите ще бъде използван паралелният анализ на Дж. Хорн. В търсене на свидетелства за наличието на един доминиращ фактор ще бъдат анализирани собствените стойности на извлечените фактори, ще бъдат приложени и подходящи графични техники. Като доказателство за едномерна структура ще се разглежда и наличието на значима разлика в собствените стойности на първия и втория фактор. За изясняване на факторната структура ще бъдат анализирани факторните тегла на отделните променливи. Важен резултат от това изследване ще бъде извличането на първоначална информация за (вероятния) модел на латентната факторна структура на ТОП на рав-

нище субтест и цялостен тест, която ще бъде тествана чрез потвърдителен факторен анализ.

Следващите три изследвания са свързани с разбирането на приложимостта като проява на очакваните свойства на теоретичния модел в емпиричните тестови данни. На оценка ще бъде подложена инвариантността (стабилността) на статистиките на въпросите от двата модела, независимостта им в рамките на един и същи теоретичен модел, както и съгласуваността между едноименните статистики от двата модела (изследователски въпроси 3, 4 и 5).

Изследване 3. Анализ на инвариантността на статистиките на тестовите въпроси

На оценка ще бъде подложена инвариантността на статистиките на тестовите въпроси, разглеждана като независимост на техните стойности от извадката от индивиди, въз основа на която са определени. В този смисъл инвариантността може да се разглежда като стабилност на статистиките.

Психометричните тестови теории, които са обект на изследването, се различават рязко по отношение на тази характеристика. В рамките на CTT индексите на въпросите са зависими от извадката от изпитани лица, т.е. при многократно оценяване на индексите на един въпрос с различни групи от изпитани е възможно да се получат различни стойности на тези индекси. При IRT оценката на параметрите на въпросите е инвариантна, независима от извадката от изпитани лица, т.е. при многократна оценка на параметрите на един въпрос с различни групи от изпитани се очаква да бъдат получени едни и същи стойности.

Тази характеристика ще бъде изследвана чрез съпоставяне на статистиките на въпросите, изчислени върху различни извадки от и. л. За тази цел ще бъде подбрана група от тестови въпроси, които са използвани в два различни тестови варианта, през различни тестови сесии. Двете извадки са непресичащи се и и. л. в тях са работили при различни условия по отношение на теста, при различна подредба на тестовите въпроси. По-конкретно, ще бъдат направени следните анализи.

Стабилност на индексите на въпросите в рамките на CTT:

3.1. Стабилност на класическия индекс на дискриминативна сила на въпросите D по CTT, определен в различни тестови условия

3.2. Стабилност на бисериалния коефициент на корелация r_{bis}, определен в различни тестови условия

3.3. Стабилност на индекса на трудност на въпросите p по CTT, определен в различни тестови условия

Стабилност на параметрите на въпросите в рамките на IRT:

3.4. Стабилност на параметъра на дискриминативната сила на въпросите a по IRT, определен в различни тестови условия

3.5. Стабилност на параметъра на трудност на въпросите b по IRT, определен в различни тестови условия

3.6. Стабилност на параметъра на налучкване на правилния отговор c по IRT, определен в различни тестови условия

Стабилността на индексите, определени в рамките на CTT, ще бъде оценена чрез прилагане на подходящи непараметрични и параметрични коефициенти на корелация. Поради ранговия характер на индексите, за прилагане на параметричната мярка техните стойности ще бъдат приведени към z-единици на единичното нормално разпределение. Чрез използване на параметричен коефициент на корелация ще бъде оценена стабилността на съответните параметри, определени в рамките на IRT. Ще бъдат анализирани и графиките на съответните корелации и при наличие на нелинейни връзки ще бъдат приложени по-адекватни методи (непараметрични корелации или методи за оценка на нелинейни корелации като коефициента η). Като допълнителна верифицираща методика ще бъдат приложени непараметрични и параметрични тестове за оценка на стабилността на статистиките.

Изследване 4. Анализ на взаимовръзките между разноименните индекси/ параметри в рамките на един и същи теоретичен модел

Изследването е предназначено да се установи дали статистиките на тестовите въпроси в рамките на даден теоретичен модел функционират самостоятелно или между тях се наблюдават определен тип взаимовръзки. За целта ще бъдат изследвани корелационните отношения от линеен тип между съответните индекси и параметри, определени въз основа на един и същи тестов вариант, при една и съща извадка от и. л. При наличие на нелинейна връзка ще бъдат приложени нелинейни регресионни методи. Ще бъдат направени и оценки на параметрите на съответните нелинейни функции. По-конкретно, ще бъдат направени следните анализи.

Изследване на взаимовръзките между индексите на въпросите в рамките на CTT:

4.1. Изследване на взаимовръзките между класическия индекс на дискриминативна сила D и индекса на трудност p

4.2. Изследване на взаимовръзките между класическия индекс на дискриминативна сила D и бисериалния коефициент на корелация r_{bis}

4.3. Изследване на взаимовръзките между индекса на трудност p и бисериалния коефициент на корелация r_{bis}

Изследване на взаимовръзките между параметрите на въпросите в рамките на IRT:

4.4. Изследване на взаимовръзките между параметрите на дискриминативна сила a и на трудност b

4.5. Изследване на взаимовръзките между параметрите на дискриминативна си-

ла a и на налучкване на коректния отговор c

4.6. Изследване на взаимовръзките между параметрите на трудност b и на налучкване на коректния отговор c.

Изследване 5. Анализ на съгласуваността между статистиките на въпросите, определени в рамките на CTT, и кореспондиращите им статистики в рамките на IRT

Изследването на съгласуваността между съответните индекси и параметри е може би най-важният аспект от съпоставителния анализ на очакваните характеристики на двете психометрични теории. Анализът ще бъде направен чрез определяне на стойностите на съответстващите си статистики по CTT и IRT върху един и същи тестов вариант, при една и съща извадка от и. л. Като основен метод за оценка на степента на съгласуваност ще бъде използван корелационният анализ и по-точно Пиърсъновия коефициент на корелация. Ще бъдат анализирани графиките на изследваните корелации и при установяване на взаимовръзки с нелинеен характер ще бъдат приложени подходящи нелинейни мерки като корелационното отношение ета (η). Крайната цел е да се определи типа на функцията, свързваща двете серии с резултати.

По-конкретно, ще бъдат направени следните анализи.

5.1. Съгласуваност на оценките на трудността на въпросите p по CTT и b по IRT

5.2. Съгласуваност на оценките на дискриминативната сила на въпросите D по CTT и a по IRT

5.3. Съгласуваност на бисериалния коефициент на корелация r_{bis} по CTT и параметъра на дискриминативна сила a по IRT

III. Резултати

Първа част: Съпоставително изследване на съответствието между допусканията на теоретичните модели и характеристиките на тестовите данни

Изследване 1. Анализ на формата на разпределенията на латентните променливи

1.1. Цели на изследването

С статистическия, а и в психометричния анализ често пъти е необходимо да се изясни каква е формата на разпределението на съответната променлива на популационно равнище. За определяне на тази форма е необходимо да се намери съответствието между емпиричното разпределение на тази променлива и някое от теоретичните разпределения, обикновено чрез съпоставяне на определени характеристики на емпиричното и съответните характеристики на избрано теоретично разпределение.

Едно от основните допускания, обичайно предшестващи статистическите анализи на данни, е за нормалност на разпределението на изследваните променливи. Това допускане, по отношение на способностите на индивидите, е едно от няколкото основните допускания, споделени от двете психометрични теории. Неговото удовлетворяване е необходимо условие за по-нататъшния психометричен анализ, тъй като почти всички негови аспекти са базирани на предпоставката, че способностите на индивидите са нормално разпределени.

Нормалното (Гаусово) разпределение е може би най-важното теоретично статистическо разпределение. Централната му роля в статистическите изследвания се обуславя от множество фактори, не на последно място от позицията му във фундаменталната централната гранична теорема (Калинов, 2002). Поради това то се използва за модел на голяма част от данните, получени по емпиричен път и следователно има огромно теоретично и практическо значение.

Макар че нормалното разпределение има дълга предистория, неговото въвеждане в областта на психологическите изследвания се дължи на теоретичните разработки на Л. Л. Търстоун през 20-те години на миналия век и по-специално работата му върху закона за сравнителните съждения (Thurstone, 1927). За да моделира разпределението на процесите на различаване за даден стимул (физичен или нефизичен) върху психологическия континуум, Л. Л. Търстоун използва нормалното Гаусово разпределение. След неговите разработки нормалното разпределение става особено популярно сред психолозите за моделиране на психологическите променливи, макар то да им

е известно и преди това (Michell, 1999).

Вярването в универсалния характер на нормалното разпределение се засилва от разработките на Р. Фишер, особено от влиятелната му книга „Statistical methods for research workers", публикувана за пръв път през 1925 г. Авторът отбелязва, че полезността на една статистика и формата на нейното разпределение се определят от формата на разпределението на признака в популацията. Ако тази форма е известна, е възможно да се определи разпределението на дадена статистика, извлечена от извадки с всякакъв обем. Приложението на тези методи се улеснява от обстоятелството, че разпределенията на много извадкови статистики се стремят към нормалното при увеличаване на обема на извадката и поради това е обичайно да се приема, че тези статистики са нормално разпределени (Fisher, 1925). Въпреки някои критики към книгата, особено за това че тя „…дава малко информация за собственото отношение на Фишер към допускането за нормалност, което формира основата на неговата работа" (Lehmann, 2008, стр. 118), идеята за „универсалната нормалност" на разпределенията пронизва целия й текст. Благодарение на Р. Фишер, който „установява нова парадигма и революционизира статистическата методология" (ibid.), на неговия огромен авторитет, тази парадигма става доминираща през следващите десетилетия.

Същевременно в голяма част от статистическите анализи нормалността на изследваните променливи негласно, но охотно се приема за даденост, без да се търсят емпирични доказателства. Както сочат резултатите от някои метаанализи, тази нагласа е изразена още по-отчетливо в полето на поведенческите и социалните науки, в които е обичайно изучаваните индивидуални характеристики да се разглеждат *a priori* като нормално разпределени. От друга страна, поради ключовата роля на нормалното разпределение, нарушаването на допускането за нормалност може да доведе до ненадеждни и дори невалидни резултати. Поради това допускането за нормалност на разпределенията на способностите следва да бъде обстойно проверено, при това върху по-широка емпирична база. Тук е уместно да си припомним двата крайни възгледа за статуса на нормалното разпределение, изразени по един блестящ афористичен начин, според единия от които „…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), а според другия „Нормалността е мит; никога не е имало и никога няма да има нормално разпределение." (Geary, 1947, стр. 241). Целта, която си поставяме, е да определим къде, между тези две (без)крайности, можем да локализираме разпределенията на способностите, измерени чрез ТОП.

1.2. Хипотези

На верификация ще бъде подложена серия от нулеви хипотези с общ вид:

H_0 ($Var\ n_{s/t}$): Съвкупността от кандидат-студенти, явили се на ТОП, са част от пе-

риодна генерална съвкупност, в които разпределението на способностите е нормално,

срещу съответните алтернативни хипотези:

H_1 ($Var\ n_{s/t}$): Съвкупността от кандидат-студенти, явили се на ТОП, са част от генерална съвкупност, в които разпределението на способностите не е нормално,

където:

$Var\ n \in n$ – пореден номер на тестов вариант (тестова батерия)

$s/t \in S/T$ - пореден номер на субтест или цялостен тест

Нулевите хипотези ще бъдат проверена при ниво на значимост α = 0.05.

1.3. Данни

Проверката за нормалност ще бъде извършена върху субтестовите и тестовите балове на 12 от тестовите варианти, подбрани за анализ. Поради обстоятелството, че всеки тестов вариант включва 11 променливи (10 субтеста и един общ тестов бал), общият брой на (суб)тестовите балове, съответно на хипотезите, които ще бъдат подложени на проверка, е 132.

1.4. Методология

Разработени са редица статистически методи за проверка на формата на емпиричните разпределения, в частност за проверка на тяхната съгласуваност с нормалното Гаусово разпределение. В своето многообразие те могат да бъдат класифицирани в четири групи въз основа на два дихотомични признака: графични – числени и дескриптивни – основани на теоретичен модел.

Графичните дескриптивни методи като хистограми, диаграми от вида „стъбло и листо" (*stem-and-leaf*) и „кутия" (*box plot*), както и тези, основани на теоретичен модел като нормална вероятностна графика (*normal probability plot*) и нормална P-P или Q-Q графика (*normal P-P, Q-Q plot*) предполагат визуална преценка на формата на емпиричното разпределение и на близостта му до теоретичното нормално Гаусово разпределение. В някои случаи графичните методи могат да бъдат резултатни, но като цяло се отличават с известна субективност на преценката. Числените дескриптивни методи, които включват определяне на индексите за асиметрия и ексцес на наблюдаваното разпределение, както и теоретично обоснованите тестове за проверка на неговата нормалност, имат предимството да бъдат по-обективни, но същевременно и те не са лишени от недостатъци. Някои изследователи поддържат мнението, че тестовете имат слаба чувствителност (мощност) при малки извадки и са свръхчувствителни при извадки с голям обем, при които дори и слаби отклонения от нормалното Гаусово разпределение се оценяват като значими (Pollard et al., 2009). Поради това би било добре обективните числени методи, най-вече специфичните тестове за проверка на хипотези за нормалност, да се използват съвместно с анализа на графичните репрезентации на разпределенията.

1.4.1. Тестове за нормалност

Статистическите тестове за нормалност спадат към групата на тестовете на съгласието (*goodness-of-fit tests*), предназначени за оценка на това дали наблюдаваното разпределение се съгласува с някое теоретично разпределение (Калинов, 2010). Доколкото в общия случай извадковите разпределения трудно постигат идеалните форми на теоретичните разпределения, тестовете на съгласието служат за оценка на това в каква степен емпиричното разпределение се „отдалечава" от теоретичното разпределение, избрано за негов модел.

Формално проблемът за съгласието може да бъде поставен по следния начин: при дадена случайна извадка $x_1, x_2, x_3, \ldots, x_n$ да се провери нулевата хипотеза, че извадката е извлечена от популация със специфична функция на разпределението $F(x)$ (Stephens, 1974). Тази задача може да бъде решена чрез различни статистически подходи, базирани на различни характеристики на нормалното разпределение. Съобразно тези характеристики тестовете за нормалност могат да бъдат обособени в три основни групи.

Тестовете, базирани на асиметрията и ексцеса, се основават на съпоставяне на тези характеристики при емпиричните и теоретичното нормално разпределение. Сред най-добрите представители на тази група тестове е комбинираният тест на Д'Агостино-Пиърсън (*D'Agostino-Pearson omnibus test*), който оценява разликите между асиметрията и ексцеса на емпиричното разпределение и съответните параметри на Гаусовото разпределение и определя единствена стойност на p като сума от квадратите на тези отклонения. Тестът комбинира два предходни теста на съгласието на Д'Агостино, базирани поотделно на оценка на отклоненията на асиметрията и на ексцеса. При справедливост на нулевата хипотеза за нормалност, тестовата статистика K^2 се приближава към разпределение χ^2 с две степени на свобода. Едно от предимствата на този тест е, че не се повлиява, ако данните съдържат съвпадащи стойности (D'Agostino, 1971; D'Agostino et al., 1990).

Една широка група от тестове на съгласието са базирани на т. н. *EDF*-статистики, чрез които емпиричната функция на разпределението $F_n(x)$ (*empirical distribution function*, *EDF*) се съпоставя с кумулативната функция на разпределението $F(x)$ (*cumulative distribution function, CDF*). Сред най-популярните са тестът на Колмогоров-Смирнов (K-S), неговата модификация, предложена от Х. Лилиефорс (*Lilliefors test for normality*), тестът на Андерсън-Дарлинг (*Anderson-Darling test for normality*), който също се разглежда като модификация на K-S, и др.

На съвършено различен принцип са построени регресионните и корелационни тестове, фокусирани върху наредените статистики. Регресионните тестове, сред които най-популярен е тестът на С. Шапиро и М. Уилк (*Shapiro-Wilk's test for normality*), се интересуват от наклона на регресионната линия, формирана от наредените статистики на

емпиричната извадка и техните очаквани стойности от теоретичното нормално разпределение, докато корелационните тестове акцентират върху силата на взаимовръзката между тях.

1.4.2. Избор на тест за нормалност

При наличието на значително разнообразие от статистически тестове за проверка на хипотезата за нормалност изборът на тест, подходящ за конкретния тип данни, следва да се основава на следните критерии: (1) мощност на теста, (2) рестрикции по отношение на обем на извадката, в която той „работи" най-добре и (3) чувствителност на тестовата статистика към особености в емпиричните данни. Нека да разгледаме как се представят тестовете за нормалност в различни ситуации.

В свое компаративно изследване М. Стефънс съпоставя качествата на 5 теста за нормалност (Колмогоров-Смирнов, Cramer-von Mises, Kuiper, Watson и Anderson-Darling), както и случаите на тяхното използване (Stephens, 1974). Интересен за нас е случай 3, в който $F(x)$ е нормално разпределение, μ и σ^2 са неизвестни и се оценяват въз основа на емпиричните данни. Сравнявайки мощността на отделните EDF-статистики с други типове тестове, авторът отбелязва, че всички те се представят отлично и срещат сериозен съперник единствено в лицето на W на Шапиро-Уилк, показвайки висока корелация с тази статистика (Stephens, 1974).

В друго компаративно изследване се съпоставят качествата на тестовете на Шапиро-Уилк и комбинираният тест на Д'Агостино-Пиърсън (Theune, 1973). Изследването е направено върху множество случайни извадки, извлечени от 4 различни разпределения, включително и нормално. При по-малките извадки ($n < 50$) мощността на Шапиро-Уилк е значително по-добра, но за средните извадки ($n \approx 100$) същото може да се каже само за някои разпределения, докато за нормалното чувствителността на двата теста е изравнена (Theune, 1973).

С. Шапиро и М. Уилк, в съавторство с Х. Чен, също провеждат широко компаративно изследване на множество тестове за нормалност, в което доказват значителната мощност на разработения от тях тест (Shapiro, Wilk, & Chen, 1968; Ryan & Joiner, 1976). Класическият тест на Шапиро-Уилк е подходящ за малки ($n < 50$) и средни по обем извадки. В няколко свои разработки Дж. Ройстън предлага модификация на тестовата статистика W, подходяща за големи извадки в диапазона $3 \leq n \leq 2000$ случая (Royston, 1982, 1986, 1989, 1992), което прави уместно използването ѝ в настоящото изследване, в което обемът на извадките е диапазона $636 \leq n \leq 1019$ и. л.

Мащабно съпоставително изследване на тестовете за нормалност предприема и Е. Сайер, обосновавайки необходимостта от него с появата на все нови и нови подходи към тази проблематика (Seier, 2002). Фактът, че съществува толкова широка палитра от подходи и конкретни тестове за проверка на нормалността говори, че нито един от тях не се радва на явно превъзходство над останалите – нещо, което и самата ав-

торка показва в своето изследване. То е основано на 50 000 извадки с малък (n = 20), среден (n = 50) и голям (n = 100) обем, симулирани по метода *Monte Carlo* от над 20 симетрични, умерено и крайно асиметрични, бимодални и балансирани смесени нормални разпределения. Авторката съпоставя 10 теста за нормалност, между които D на Колмогоров-Смирнов, A^{2*} на Андерсон-Дарлинг, Y на Д'Агостино, W на Шапиро-Уилк, z на Ройстън и др., които са представители на трите групи тестове, по редица показатели като мощност, леснота за изчисляване, валидност на статистическата значимост p, информацията, която предоставят и дори по наличността им в статистическите софтуерни пакети.

Резултатите от изследването сочат, че по отношение на емпиричните стойности на грешките от първи род (при фиксирани стойности на α = 0.01, 0.05 и 0.10) за симетрични разпределения, различните тестове имат едно и също поведение, независимо от обема на извадката. Получените емпирични стойности на α се отклоняват съвсем слабо от фиксираните стойности.

Аналогични са резултатите и при проверката на мощността на тестовете върху симетрични разпределения при α = 0.05. Преимуществото на една или друга група тестове зависи от типа на разпределението или от обема на извадката. Така например регресионните тестове имат по-висока мощност в сравнение с останалите при симетрични бимодални разпределения, но при разпределения с висок ексцес преимущество имат тестовете, основани на асиметрията и ексцеса. Анализът на резултатите от изследването на мощността на тестовете при балансирани смесени нормални разпределения, както и при асиметрични разпределения, води към същите изводи.

Все пак предпочитанията на авторката, особено в случаите, в които целта на тестването е да се определи дали едно разпределение е нормално или не, определено клонят към регресионните тестове, които се характеризират системно с по-висока мощност, демонстрирана върху широка гама от разпределения. Сред тях се открояват тестът QH^* на Чен-Шапиро, както и W на Шапиро-Уилк (Seier, 2002; Chen & Shapiro, 1995).

Въз основа на анализираната по-горе информация за качествата на отделните тестове, както и на критериите за подбор, при проверката на нулевите хипотези ще бъдат използвани два статистически теста, принадлежащи към различни категории: класическият тест на Колмогоров-Смирнов (с използване вероятностите на Лилиефорс) и тестът на Шапиро-Уилк, като водещ при вземането на решения относно нулевите хипотези ще бъде тестът на Шапиро-Уилк, който отговаря най-добре на формулираните по-горе изисквания.

Тестове на Колмогоров-Смирнов и Шапиро-Уилк

Едноизвадковият тест на Колмогоров-Смирнов е сред най-широко използваните тестове на съгласие, основан на сравнението (максималната положителна или отрица-

телна разлика) между емпиричното кумулативно разпределение и очакваното (теоретично) кумулативно нормално разпределение. Особеност на класическия тест на Колмогоров-Смирнов е, че асоциираната с тестовата статистика вероятност p се изчислява въз основа на параметрите на нормалното разпределение (μ, σ), които следва да бъдат предварително известни. В повечето случаи обаче стойностите на тези параметри са неизвестни и се изчисляват въз основа на емпиричните данни. В тези случаи следва да се прилагат т. н. статистики на Лилиефорс, описващи вероятността от получаване на определена (или по-голяма) стойност на проверяващата статистика D в зависимост от μ и σ, оценени чрез конкретните данни (Lilliefors, 1967; Molin & Abdi, 1998).

Тестът на Лилиефорс се разглежда като модификация на Колмогоров-Смирнов и очевидно е полезен в случай 3. на М. Стефънс, в който разпределението, фиксирано в нулевата хипотеза, е определено, но с неизвестни параметри, каквото е разпределението на способностите в генералната съвкупност. Мощността на теста, макар и не толкова висока, колкото на останалите, е съпоставима с тях. Няма литературни данни за рестрикции по отношение на обема на извадката или за повлияване от особености на данните. За двустранен тест проверяващата статистика D_n се изчислява по формулата:

$$D_n = \sup_{-\infty<x<+\infty} \left| F_n(x) - F(x) \right| \tag{58}$$

където:

$F_n(x)$ - емпирична кумулативна функция на разпределението

$F(x)$ - нормална кумулативна функция на разпределението с параметри, оценени чрез извадката.

Тестът за нормалност на Шапиро-Уилк е сред най-дискутираните и използвани тестове за нормалност на разпределението преди всичко заради високата си чувствителност в сравнение с алтернативните тестове (Shapiro & Wilk, 1965; Theune, 1973; Stephens, 1974). Той се базира на разпределението на наредените статистики *(order statistics*) и неговата проверяваща статистика W се изчислява по следната формула:

$$W = \frac{1}{\sum_{i=1}^{n} \left(x_i - \overline{X} \right)^2} \left[\sum_{i=1}^{k} a_i \left(X_{(n-i+1)} - X_{(i)} \right) \right]^2 \tag{59}$$

където:

n - обем на извадката

$\overline{X}$ - извадкова средна стойност

$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ - емпирични стойности, подредени във възходящ ред

X_i- i-та наредена статистика

$a_1, a_2, \ldots, a_k$ - константи (таблични стойности), образувани от средните, дисперсиите и ковариациите на наредените статистики в извадка с обем n, извлечена от нормално разпределение

k - приблизително равно на $n/2$

Тестовата статистика може да се изменя в интервала $0.00 < W \leq 1.00$. Високи стойности на W, близки до 1.00, са индикация за нормалност на изследваното разпределение. При извадки с много малък обем мощността на теста би могла да намалее, а при много големи извадки – неговата точност. Друг негов недостатък е, че се повлиява от повтарящи се стойности, но в по-малка степен, отколкото тестът на Anderson-Darling, с който най-често е сравняван.

В съгласие с идеите на комплексния изследователски анализ на данни (EDA), обоснован от Дж. Тюки (Tukey, 1977; Hartwig & Dearing, 1979; Behrens, 1997), в настоящото изследване, освен числените методи, основани на теоретичен модел, ще бъдат приложени и елементи от останалите три подхода. За всяко емпирично разпределение ще бъдат определени и стойностите за асиметрия и ексцес, ще бъдат приложени и подходящи графични методи.

1.5. Резултати

В рамките на верификацията на формулираните статистически хипотези са проверени разпределенията на компонентите на избраните 12 тестови батерии (на ниво субтестов и общ тестов бал), като за всеки компонент са изчислени следните 8 статистики: Колмогоров-Смирнов D, Lillefors p, Shapiro-Wilk's W и p, асиметрия, стандартна грешка на асиметрията, ексцес и стандартна грешка на ексцеса. Общият брой на проверените разпределения е 12 X 11 = 132, а на изчислените статистики – 1,056. Получените резултати са представени в приложение 3.

Приоритетен сред приложените тестове за проверка на нормалността на изследваните разпределения е тестът на Шапиро-Уилк. При последователната проверка на хипотезите за тяхната форма проверяващата статистика W варира в сравнително тесните граници от 0.907 до 0.997. Тъй като високите стойности на проверяващата статистика W, клонящи към 1.00, са индикация за нормалност на съответното разпределение, бихме могли да очакваме, че една значителна част емпиричните разпределения на тестовия бал се съгласуват с Гаусовото.

Противно на наложилото се мнение за нормалното разпределение като „златен стандарт", резултатите от направените проверки показват обратното. В 129 от разгледаните случаи (97.73% от всички изследвани разпределения) резултатите водят до отхвърляне на съответната нулева хипотеза, тъй като проверяващата статистика W е значима при равнище p = 0.00, в 1 случай (0.76%) – на равнище p = 0.02 и в още един случай (0.76%) – на равнище p = 0.04. Тестовата статистика W не е значима само при

едно разпределение – при общия бал на вариант 166, при който равнището на нейната значимост $p = 0.16$.

По подобен начин изглеждат резултатите от проверката на хипотезите за нормалност с използване на теста на Лилиефорс, чиято проверяваща статистика D се изменя в границите от 0.034 до 0.204. В 131 от случаите (99.24% от всички изследвани разпределения) тя е значима на равнище $p < 0.01$ и в 1 от тях (0.76%) – на равнище $p < 0.05$. При нито едно от разпределенията не се появяват основания за разглеждане на съответната нулева хипотеза като справедлива. Следователно успоредното прилагане на двата статистически теста за нормалност, представляващи различни подходи за проверка на нулевите хипотези, води до напълно съгласувани резултати, които се различават драматично от общоприетите схващания за формата на разпределенията в психометричните изследвания.

Нека да разгледаме разпределенията на тестовите балове и от друг ъгъл, като изчислим емпиричните стойности на асиметрията и ексцеса на изследваните разпределения. Това са две описателни статистики, които също носят важна информация за тяхната форма и по-конкретно за степента, в която се отклоняват от нормалното.

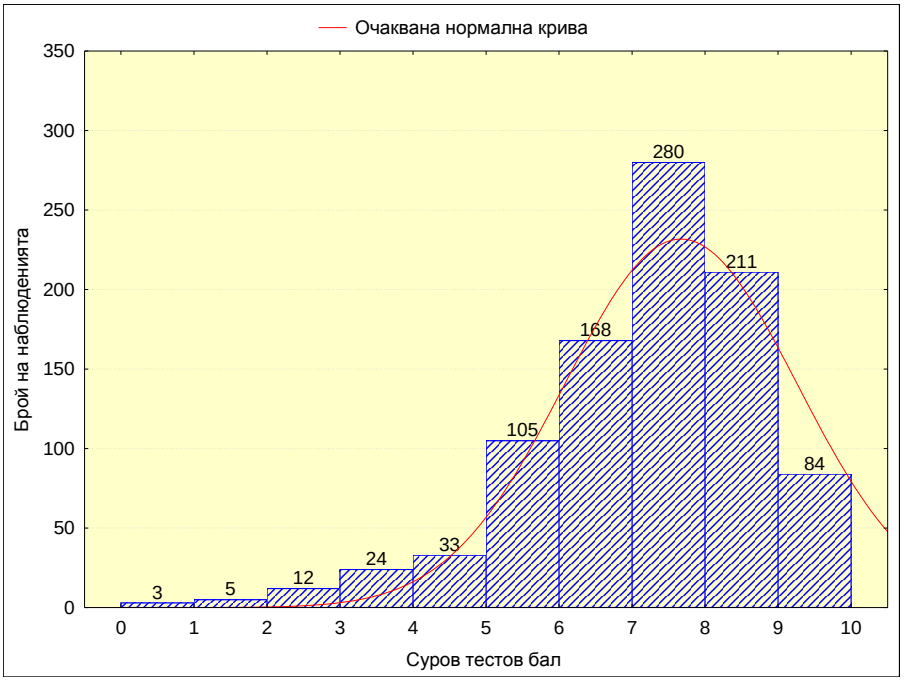
Индексите за асиметричност на разпределенията се изменят в границите от -1.024 до 0.796, като по-голяма част от тях имат положителна асиметрия. Такива са 78.03% от всички анализирани разпределения, като най-често срещаните (модални) стойности са в интервала 0.20; 0.40, в който са разположени 32.58% от разпределенията. Графиката от фигура 1, представена в приложение 4, демонстрира изместването на основната маса от разпределения вдясно от нулевата стойност на индекса за асиметрия, отличителен за нормалното разпределение. Като цяло, със системно по-високи стойности на асиметрия се характеризират тестовите раздели от цикъла на природо-математическите дисциплини – 5. *Математика*, 6. *Физика*, 7. *Химия* и 8. *Биология*, а с по-ниски – 1. *Български език* и 10. *Семантика*.

Индексите за ексцес на разпределенията варират от -0.700 до 1.539, като преобладаващата част се характеризират с отрицателен ексцес. Такива са 63.64% от анализираните разпределения, като най-висок е техния дял в интервала -0.20; 0.00, който включва 26.52% от всички разпределения. Фигура 2, представена също в приложение 4, показва тенденцията за отместването на стойностите на този индекс наляво от нулевата стойност, характерна за нормалното разпределение. Разделите с преобладаващо високи индекси на ексцес са отново от цикъла на природо-математическите дисциплини – 5. *Математика*, 6. *Физика*, 7. *Химия* и 8. *Биология*, а с по-ниски – 1. *Български език* и 9. *Разсъждения*.

Строго погледнато, нито едно от разпределенията не се отличава с нулева асиметрия или ексцес. Като примери нека да разгледаме графичните репрезентации на две разпределения, които онагледяват две „крайни”, „гранични” състояния на тестовите балове в разглежданата съвкупност. Първото е разпределението на тестовия бал

на раздел 10. Семантика от вариант 175 (n = 925), чиито проверяващи статистики D на Колмогоров-Смирнов и W на Шапиро-Уилк са значими на ниво α = 0.05 и при това се отличава от останалите с най-високи стойности на асиметрия (-1.024) и ексцес (1.539), т. е. е най-отдалечено от нормалното. Второто е разпределението на общия тестов бал от вариант 166 (n = 835), което е единственото сред избраната съвкупност, чиято проверяваща статистика W не е значима на ниво на α = 0.05 (W = 0.997; p = 0.16), т.е. едно нормално емпирично разпределение със стойности на асиметрия 0.112 и ексцес 0.054.

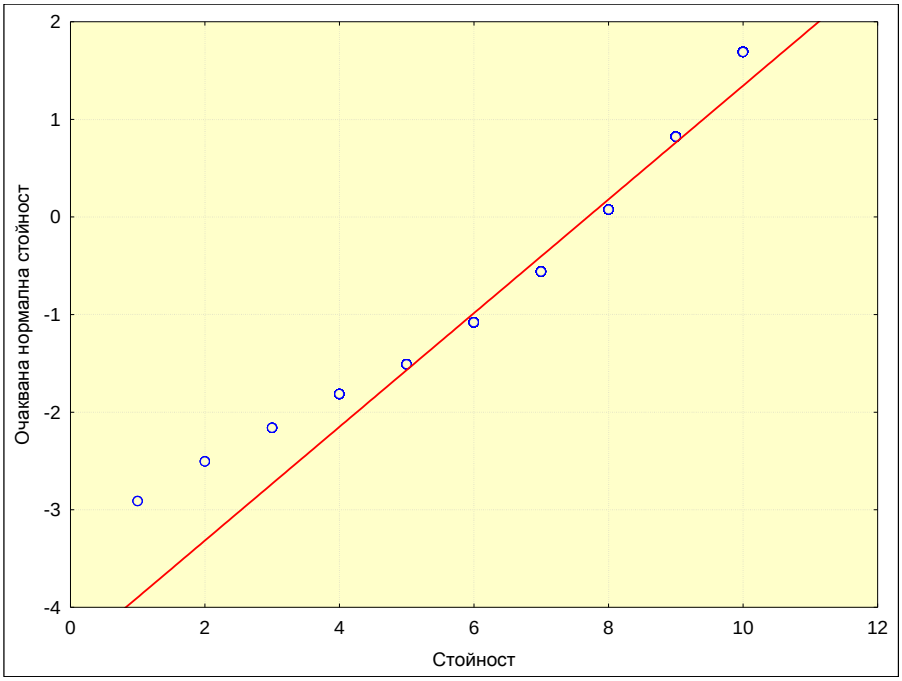
Фигура 4. Хистограма на разпределението на тестовия бал на раздел 10. Семантика от вариант 175 (n = 925)



На горната хистограма отчетливо се открояват отбелязаните особености на разпределението – неговата силно изразена асиметричност с натрупване на основната маса от и. л. в дясната половина на скалата, особено в интервала от 7 – 9 точки, и дългата „тежка” опашка в посока към ниските тестови балове в лявата половина на скалата, както и силно изразеният ексцес, особено при модалната стойност от 8 точки.

Нормалната вероятностна графика, която обикновено се използва успоредно или вместо хистограмата при анализа на нормалността на разпределението в рамките на *EDA*, разкрива характерните особености на това разпределение. Отклоненията на стойностите на анализираната променлива от правата линия са индикация за отклонението на нейното разпределение от нормалното. На графиката, представена на фигура 5, се вижда ясно изразената U-образна форма разсейването на тестовия бал на раздел 10. Семантика от вариант 175.

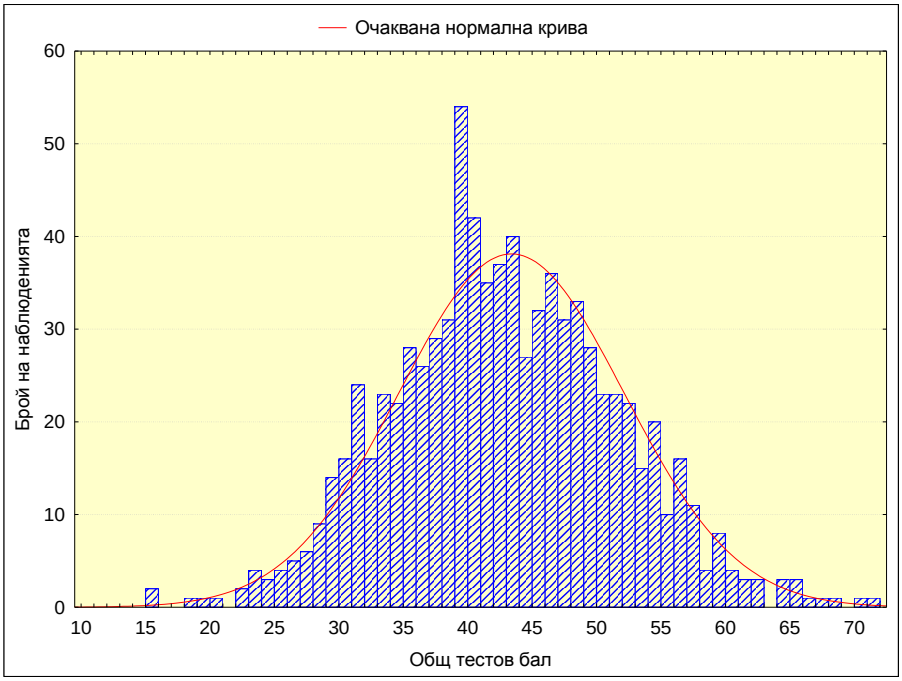
Фигура 5. Нормална вероятностна графика на разпределението на тестовия бал на раздел 10. Семантика от вариант 175 (n = 925)



Интерпретацията, която може да бъде направена съгласно подхода на Тюки и неговите колеги (Hoaglin, Mosteller & Tukey, 1991, стр. 187), е за наличието на значителна негативна асиметрия, като отклоненията от формата на нормалното разпределение са по-ясно изразени в левия край на разпределението.

На следващата графика е представена хистограмата на разпределението на общия тестов бал от вариант 166 (n = 835), което, съгласно резултатите от теста на Шапиро-Уилк, може да бъде разглеждано като апроксимация на нормалното.

Фигура 6. Хистограма на разпределението на общия тестов бал от вариант 166

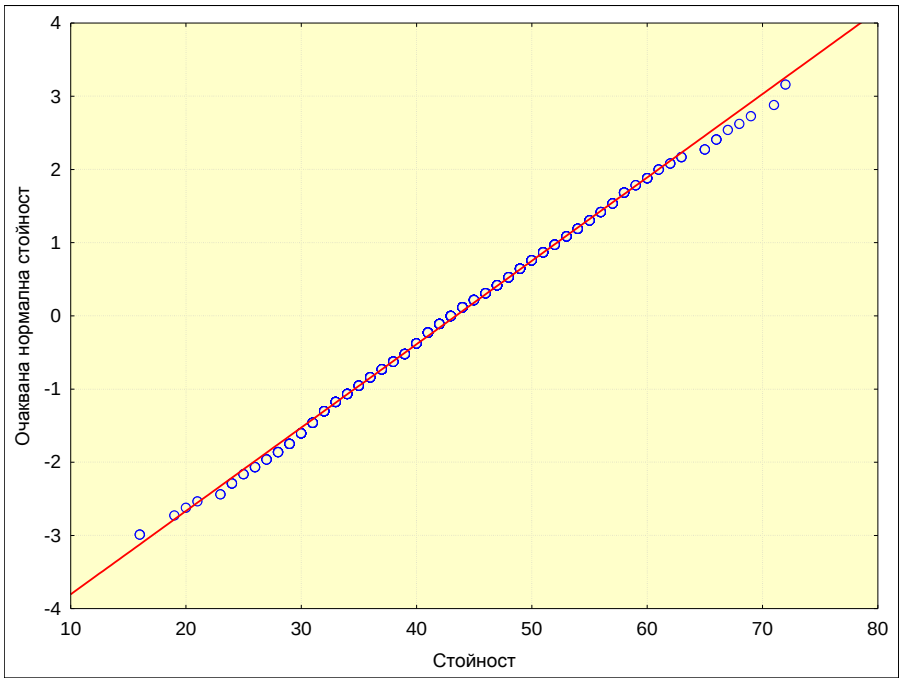


Ниските стойности на индексите за асиметрия и ексцес на разпределението от горната хистограма се съгласуват с почти идеалната му форма. Въпреки това, като не забравяме, че то отразява емпирични данни, бихме могли да отбележим повсеместния „излишък” и „недостиг” на честоти в почти всяка негова точка.

Нормална вероятностна графика на същото разпределение, показана на Фигура 7 показва почти линейната подредба на стойностите на разглежданата променлива. Те имат слабо различаваща се S-образна форма и се отклоняват по-видимо от правата линия в двата й края, като левият й край е над линията, а десния – под нея, което, според Тюки и неговите колеги, следва да се интерпретират като знак за по-„леки” опашки на наблюдаваното разпределение. Отклонения, също така слаби, се забелязват и във вътрешната зона на правата, свидетелстващи за по-ниски честоти в областта на 24 – 30 точки, както и за по-високите честоти в интервала 40 – 44 точки.

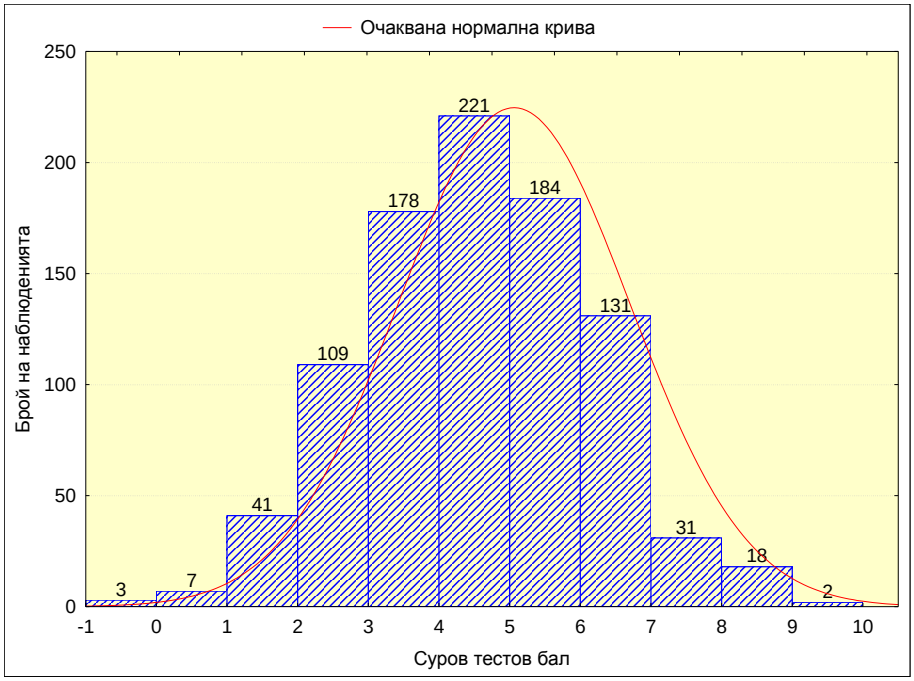
Впрочем немалко други разпределения имат съпоставими с тези на вариант 166 (n = 835) и дори по-ниски индекси на асиметрия и ексцес.

Фигура 7. Нормална вероятностна графика на разпределението на общия тестов бал от вариант 166



Такова е например разпределението на тестовия бал на раздел 8. Биология от вариант 175, представено на фигура 8 и 9 под формата на хистограма и на нормална вероятностна графика, което се характеризира с най-ниски (в сравнение с останалите разпределения) индекси на асиметрия и ексцес, които са близки до нула (съответно 0.003 и -0.088).

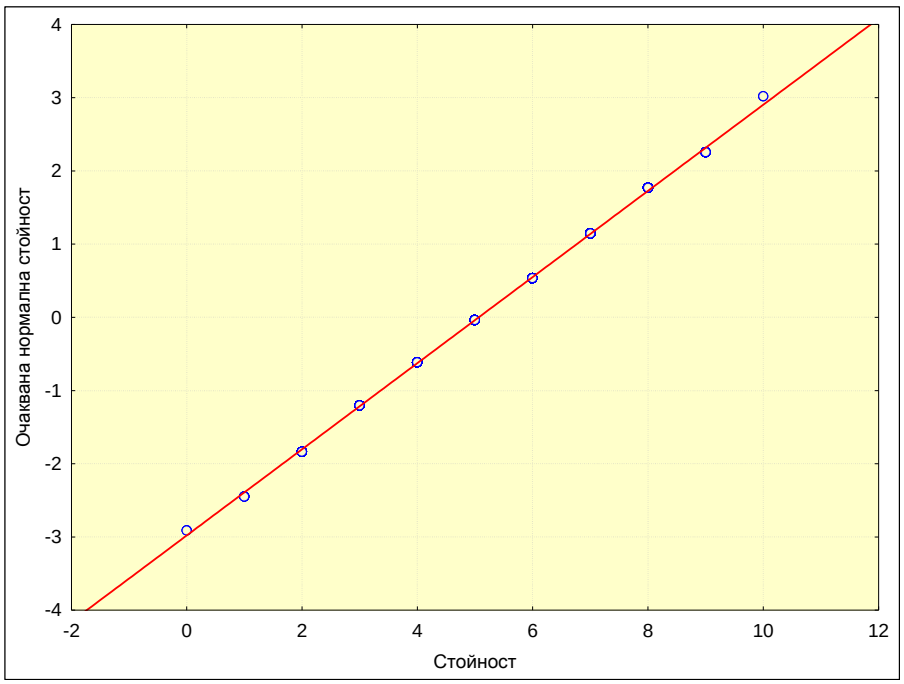
Фигура 8. Хистограма на разпределението на субтестовия бал на раздел 8. Биология от вариант 175 (n = 925)



Хистограмата на горната графика показва симетричната, почти идеална форма на разпределението на субтестовия бал.

На следващата графика се забелязват минимални отклонения на подредбата на стойностите на съответната променлива от правата линия, което може да се интерпретира като белег за нормалността на съответното разпределение.

Фигура 9. Нормална вероятностна графика на разпределението на тестовия бал на раздел 8. Биология от вариант 175 (n = 925)

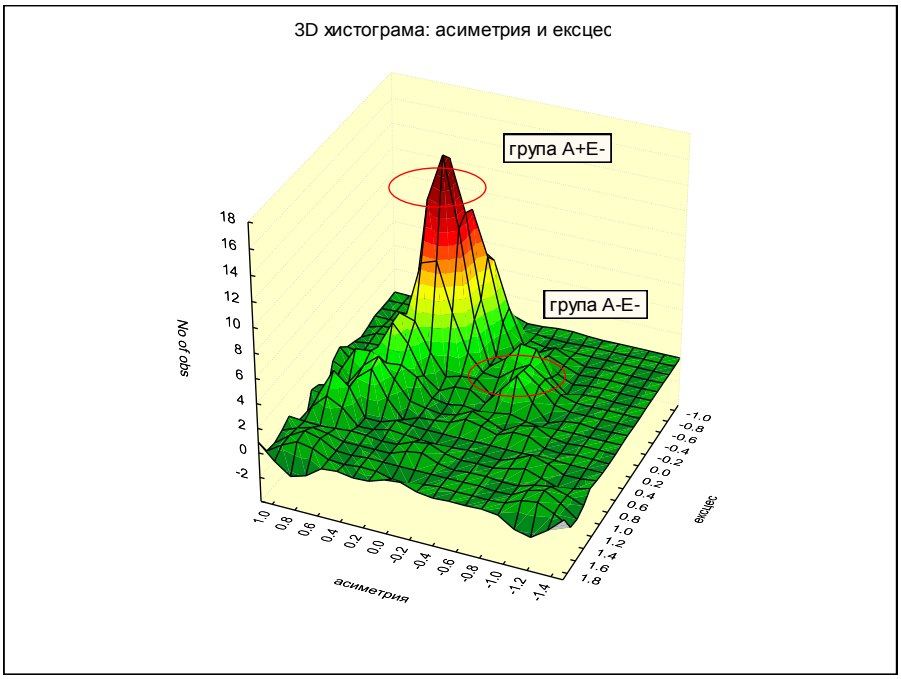


Анализът горните две графични изображения, ако не се вземе предвид предварителната информация за резултатите от тестовете за нормалност, би довел до заключението за пълно съответствие на това разпределение с Гаусовата крива, но резултатите от теста на Шапиро-Уилкс (W = 0.967, p = 0.00) водят до отхвърляне на съответната нулева хипотеза относно неговата форма.

Очевидно е, че прилагането на различни (графични и числени) методи за определяне на нормалността, при някои разпределения води до вземането на противоречиви, дори взаимноизключващи се решения. Поради това обстоятелство някои изследователи се придържат към практическото правило, съгласно което разпределение, чиито индекси на асиметрия и ексцес едновременно не надхвърлят -1.00 или +1.00, се третират като нормални. Такива са 127 от разпределенията (96.21% от всички), като индексите на асиметрия варират в интервала (-0.691; 0.783), а тези на ексцес – в интервала (-0.700; 0.965).

Поради това, че горното правило е твърде субективно, грубо и очевидно би довело до голям брой неправилни решение, би било полезно да проследим едновременното изменение на стойностите на двата индекса в отделните разпределения и дали чрез този похват може да бъде разкрита някаква закономерност, никаква отчетлива тенденция. За целта бихме могли да разгледаме разпределението на всеки субтест или тест като отделен обект (случай), който има две характеристики: асиметрия и ексцес. Двата вектора от стойности, съответно за асиметрия и ексцес, формират (условно) съвместно двумерно разпределение, представено на следващата 3D хистограма.

Фигура 10. Съвместно двумерно разпределение на асиметрията и ексцеса на разпределенията на тестовите балове



На горната фигура се забелязват отчетливо няколко характерни особености на наблюдаваните случаи. На първо място следва да отбележим неравномерното разпределение на честотите на техните стойности за асиметрия и ексцес. Наблюдават се, макар и малко на брой, тестови балове с висока асиметрия, с висок ексцес или и с високи стойности по двата индекса. От друга страна, забелязва се определено натрупване на честоти в една сравнително ограничена, но все пак широка площ върху хоризонталната основа на тримерната графика. По-точно, можем да говорим за две конфигурации от тестови балове. Едната от тях, обозначена условно като група A-E-, включва по-малък брой тестови балове, характеризиращи се предимно с отрицателна асиметрия (A-) и отрицателен ексцес (E-). Другата, обозначена като група A+E-, включва преобладаващата част от тестовите балове, характеризиращи се предимно с положителна асиметрия (A+) и отрицателен ексцес (E-). Тъй като в двете групи преобладават тестовите балове с отрицателен ексцес, можем да обобщим, че в своята съвкупност те показват устойчива тенденция да бъдат с отрицателен ексцес. По-голяма вариативност се наблюдава по отношение на тяхната асиметричност, като преобладават тестовите балове с положителна асиметрия.

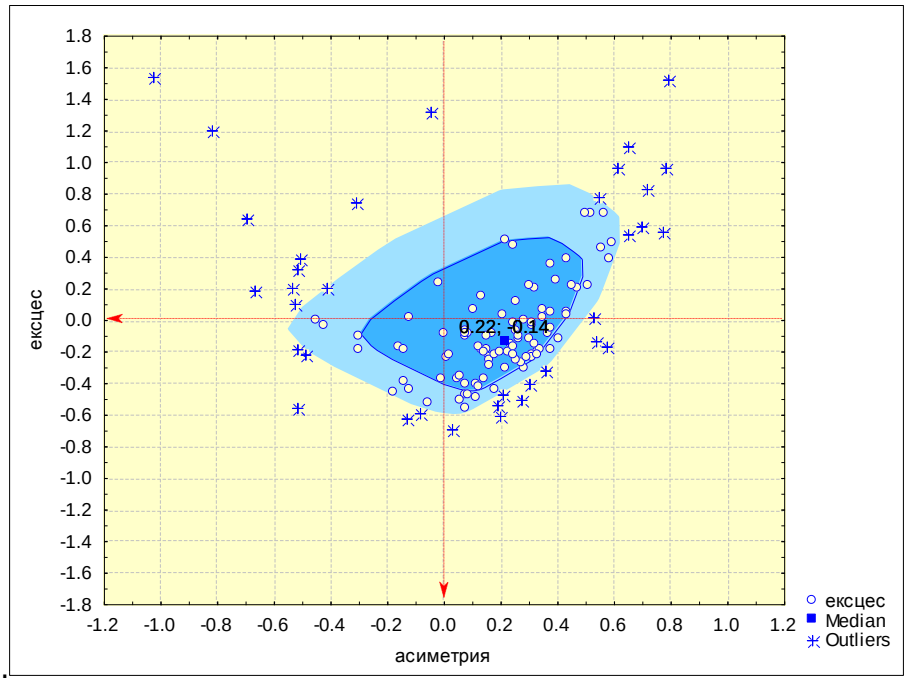
Ако потърсим една единствена оценка на отбелязаните тенденции, бихме могли да използваме графичната техника, известна като диаграма тип „чувал" (*bag plot*), разработена от Дж. Тюки (Rousseeuw, Ruts & Tukey, 1999).

Този тип диаграма, предназначена за изследване на двумерни разпределения, е развитие на аналогичната едномерна статистическа техника, основана на диаграмата тип „кутия" (или квартилна диаграма) (*box and whiskers plot*), предложена също от Дж. Тюки във фундаменталния му труд за изследователския анализ на данни (Tukey, 1977). Обикновено диаграмата тип „чувал" се разполага върху диаграмата на разсейване на стойностите на променливите, формиращи двумерното разпределение, и дава възможност за изследване на тяхната локализация и разсейване (Stryjewski, 2010). В структурно отношение графиката съдържа всички компоненти на едномерната диаграма тип „кутия", видоизменени съобразно нейното различно предназначение, както е показано на фигура 11.

Един от основните компоненти на графиката е „чувалът" (областта с тъмносин цвят), който съответства на „кутията" в едномерната графика. Този компонент характеризира разсейването на стойностите и съдържа в себе си 50% от всички данни с най-голяма дълбочина. Точките, обозначаващи отделните тестови балове, са концентрирани в долната дясна част на тази област, отразявайки тенденцията разпределенията да се характеризират с определена (положителна) асиметрия и (отрицателен) ексцес. Формата на тази област (обикновено представяна като изпъкнал полигон) на фигура 11 съответства на нейното наименование – „чувалът" е леко издължен от долния ляв към горния десен ъгъл на графиката, обозначавайки по този начин положителната връзка между двата индекса, корелацията между които обаче е слаба (r = 0.122) и статисти-

чески незначима при α = 0.05. Неговата площ е относително малка, което е свидетелство за наличието на ядро от тестови балове с относително ниска асиметричност и ексцес.

Фигура 11. Диаграма тип „чувал” на съвместното разпределение на индексите на асиметрия и ексцес



Втората област (със светлосин цвят) е ограничена от „ограда” (*fence*), която отделя „вътрешните” (*inliers*) от „външните” (*outliers*) точки. Оградата се образува чрез „надуване” на чувала от неговия център навън, с определен коефициент, който при горната графика е 1.50. Светлосинята зона е „обвивка” на чувала и съдържа сравнително малък брой тестови балове („вътрешни” точки).

Най-важният компонент в графиката е медианата на двумерното разпределение (двумерна медиана, „медиана на Тюки”), която отразява концепцията на Дж. Тюки за „дълбочината на данните” (*data depth*), по скоро за дълбочината на всяка точка, представляваща елемент от някакво двумерно разпределение. При едномерно разпределение медианата е средната точка във вариационния ред (50-тия процентил), характеризира се с максимална дълбочина и се използва като оценка на централната тенденция. Дж. Тюки разширява приложението на тази концепция за многомерни случаи (Tukey, 1975; 1977; Rousseeuw & Ruts, 1998; Rousseeuw, Ruts & Tukey, 1999).

При двумерен случай дълбочината (*halfplane location depth*) на една точка $\Theta \in \Re^2$, свързана с двумерното разпределение $X = \{x_1, x_2, \ldots, x_n\}$ се определя като най-малкия брой точки в двумерното пространство, разположени в едната му половина, определена от права, минаваща през същата точка Θ (Tukey, 1975). Медианата на Тюки е точка Θ с максимална дълбочина (*deepest depth*) k^*. Ако тази точка не е единствена, медианата на Тюки се определя като център на тежестта на множеството от точки

с дълбочина k^* (Rousseeuw & Ruts, 1996; 1998; Miller et al., 2003).

Като мярка на централната тенденция медианата (при едномерен или двумерен случай) се характеризира със статистическа устойчивост, тъй като чрез нея се елиминира влиянието на екстремните стойности (Micceri, 2003; Ripley, 2004). Медианата на Тюки от фигура 11, която е централна точка на чувала, има координати +0.22 (асиметрия) и -0.14 (ексцес), което означава, че в своята съвкупност разпределенията на тестовите балове проявяват тенденция да бъдат скосени отляво и по-плоски от нормалното.

1.6. Дискусия

Съгласно концепцията на Дж. Тюки, при анализа на формата на разпределенията на тестовите балове бяха приложени различни статистически методи, принадлежащи към четирите категории, описани по-горе. Като цяло решенията, които могат да се вземат въз основа на получените резултатите, са съгласувани, макар че при някои конкретни разпределения тези резултати водят и до противоречия.

Прегледът на изследванията, посветени на подходите за проверка на нормалността на разпределенията, показа, че нито един от тях не се отличава с категорично превъзходство над останалите. Все пак предпочитанията са към методите, основани на теоретичен модел и в частност - поради тяхната обективност, към числените методи. Разгледаните съпоставителни изследвания на статистическите тестове за нормалност, базирани на различни характеристики на нормалното разпределение, в повечето случаи определят водеща позиция на теста на Шапиро-Уилк, който в настоящото изследване бе избран за основен статистически критерий.

Резултатите от направените проверки на формулираните хипотези за нормалност показват, че проверяващата статистика W на Шапиро-Уилк е значима при равнище $\alpha = 0.05$ при огромна част от анализираните разпределения. Такива са резултатите при 131 от общо 132 разпределения на тестовите балове (99.24% от всички). В тези случаи може да се вземе решение за отхвърляне на нулевите хипотези за нормалност на разпределението на способностите в генерална съвкупност от кандидат-студенти. Сходни са резултатите от проверката на нулевите хипотези с алтернативния тест за нормалност на Лилиефорс, при която до отхвърляне на нулевите хипотези се достига при 100% от разгледаните разпределения.

Анализът на разпределенията чрез числените дескриптивни методи показва широките граници на изменение на индексите за асиметрия и ексцес, като при някои от разпределенията те достигат екстремни стойности. Отклоненията от стандартните за нормалното разпределение стойности е както в положителна, така и в отрицателна посока, с превес на разпределенията с положителна асиметрия и отрицателен ексцес.

Наличието на тенденция за преобладаване на положително асиметричните и по-плоски разпределения се потвърждава и от анализа на данните чрез графичния ме-

тод тип „чувал” на Дж. Тюки. Разгледан в по-широките граници на „чувала” от фигура 11, типичният индекс на асиметрия варира от -0.30 до +0.50, а на ексцес – от -0.42 до +0.52. Останалите 50% от тестовите разпределения попадат извън тези граници. Следователно може да се направи обобщението, че по отношение на тези два основни индикатора, отклоненията от нормалната крива, нерядко достигащи до екстремни стойности, са твърде много и са по-скоро обичайни, отколкото изключения. Координатите на двумерната медиана на Тюки, която е мярка на централната тенденция, сочат, че типичното разпределение на тестовия бал е скосено отляво и по-плоско в сравнение с нормалното.

В специализираната литература няма единно мнение относно границите на изменение на асиметрията и ексцеса на емпиричните разпределения, за да бъде разглеждана съответната променлива като нормално разпределена. По-горе беше споменато за практическото правило, съгласно което като нормални се третират разпределения, чиито индекси на асиметрия и ексцес едновременно не надхвърлят -1.00 или +1.00. Такива са над 96% от разглежданите разпределения. Ние разглеждаме обаче това правило за твърде либерално, грубо и неточно, тъй като в тези граници попадат множество разпределения, които следва да се разглеждат като отклоняващи се от нормалното не само въз основа на тестовете на съгласието, но и на дескриптивните графични методи.

В свое изследване Х. Парк изказва съвсем основателното мнение, че стойностите на двата индекса трябва да бъдат близки до 0.00. Въпреки това авторът разглежда стойностите в интервала (-0.40; 0.40) като индикация за разпределение, близко до нормалното, но все пак заключава, че тези две описателни статистики не предоставят такъв тип информация, който да бъде достатъчно убедителен за вземане на решение относно нормалността на разпределението (Park, 2008).

При по-голяма част от разпределенията, анализирани чрез графичните методи, могат да бъдат забелязани по-силни или по-слаби отклонения от Гаусовата крива. Същевременно графичните анализи на някои от разпределенията, както и числените дескриптивни методи, водят до решения, които са противоположни на тези, които следват от паралелно приложените числени методи, основани на модел. Наблюдават се разпределения с изключително ниски стойности на асиметрия (например вариант 198, раздел 9. *Разсъждения* с индекс -0.006), на ексцес (вариант 192, раздел 10. *Семантика* с индекс 0.002) или едновременно на двата индекса (вариант 175, раздел 8. *Биология* с индекси съответно 0.003 и -0.088), за които тестовете на съгласието сочат отсъствие на нормалност. Хистограмите на тези, а и на други разпределения, както и нормалните вероятностни графики, основани на модел, също биха довели до решения за нормалност на техните разпределения. Нещо повече, противно на очакванията за един образователен (суб)тест с ограничен брой категории, при нито едно от разпределенията на отделните раздели не се наблюдава ярко изразен подов или таванен

ефект.

Въз основа направените анализи могат да се направят следните обобщения:

(1) Паралелното използване на различни техники за проверка на хипотезите за нормалност на разпределенията на латентните способности в по-голяма част от случите води до съгласувани решения за отдалеченост на съответното разпределение от Гаусовата крива. В някои случаи обаче то води до противоречиви решения. Без съмнение числените методи за проверка на нормалността на разпределенията се оказват по-консервативен подход, особено статистическите тестове, които водят до почти последователно отхвърляне на всички предположения за нормалност. По-либерални са графичните методи, в частност дескриптивните, които, макар и при сравнително малък брой разпределения, водят до решения за нормалност.

(2) Свидетелствата за нормалност на разпределенията са сравнително малко на брой, а и произтичат от методи, които в настоящото изследване са с по-нисък приоритет. Ето защо, главно въз основа на резултатите от тестовете на съгласието, чрез които са проверени нулевите хипотези, може да се направи изводът, че по правило разпределенията на тестовите балове в ТОП се отклоняват от нормалното Гаусово разпределение, т. е. разпределението на способностите не е нормално. Оттук следва, че между теоретичното допускане (изискване) на тестовите теории, по-специално на Теорията за отговор на тестов въпрос, за нормалност на разпределението на латентните променливи, и съответната характеристика на наблюдаваните данни, по правило липсва съответствие. Това несъответствие поставя под въпрос приложимостта на тези теории към данните, получени чрез Теста по общообразователна подготовка.

(3) Означава ли това, че образната мисъл на Р. Гиъри „…нормалността е мит: никога не е имало и никога няма да има нормално разпределение” е универсално вярна? Ако Р. Гиъри има предвид теоретичната нормална Гаусова крива, то в този случай вероятно нито едно извадково или популационно разпределение няма нейната идеална форма. Ние обаче не можем да се съгласим с него. При цялото разнообразие от форми на наблюдаваните емпирични тестови разпределения (разпределения на различни способности), именно Гаусовата крива, а не друга, е тяхната най-подходяща теоретична апроксимация. Разпределенията се различават само по степента на отдаленост от този теоретичен модел. Както беше показано, някои от тях толкова се доближават до теоретичната крива, че минават теста за нормалност (това е единичният случай с общия бал на вариант 166), или имат нищожни индекси за асиметрия и ексцес. По-съществените отклонения от нормалната форма обаче са по-скоро правило, отколкото изключение. Или, ако трябва да перифразираме твърдението от предходния параграф, типичното разпределение на тестовия бал се отклонява от нормалното повече, отколкото е приемливо и допустимо за прилагане на избраните базови модели.

Възможни обяснения

Разгледани в контекста на традицията, завещана от Л. Л. Търстоун, който пръв използва нормалното разпределение за моделиране на психологически променливи, а и на практиката в психологическите изследвания, получените резултати изглеждат озадачаващи.

Едно възможно обяснение на това обстоятелство може да бъде потърсено в мнението на някои изследователи, според които при прилагане на тестовете за нормалност се наблюдава тенденция тестовите статистики да бъдат по-малко чувствителни при извадки с по-малък обем, които почти винаги „преминават" тестовете, и по-чувствителни при извадки с по-голям обем, което води по-често към отхвърляне на нулевата хипотеза (Park, 2008). Но би могло да се изкаже и мнение в противоположен смисъл – с нарастване на обема на извадката той се приближава до този на генералната съвкупност, а формата на нейното разпределение – до нормалното, ако генералната съвкупност също е разпределена нормално.

Данните от цитираните по-горе съпоставителни изследвания на различни тестове за нормалност на базата на симулирани извадки обаче не подкрепят подобно становище – тестовете имат стабилно поведение при извадки с различен обем. Вярно е обаче, че в по-голяма част от тези изследвания авторите боравят с извадки с обеми, значително по-малки от обемите на извадките в настоящото изследване.

В цитираното по-горе изследване на Х. Парк авторът представя резултатите от собствено изследване, в което съпоставя тестовите статистики и асоциираните с тях вероятности от 7 теста за нормалност, включително и използваните в настоящото изследване, приложени върху случайни извадки с различен обем (10, 100, 500, 1 000, 5 000 и 10 000 наблюдения), последните две от които значително надхвърлят обемите на извадките в настоящото изследване. Извадките са генерирани от стандартно нормално разпределение, симулирано чрез статистическия пакет SAS. Показателно е, че нито един от тестовете не води до отхвърляне на нулевата хипотеза, независимо от обема на извадката. Нещо повече, стойността на тестовите статистика W на Шапиро-Уилк очаквано расте с нарастване на обема на симулираните извадки (например при n = 500, W = 0.9956, p = 0.1680, а при n = 1 000, W = 0.9980, p = 0.2797), докато D на Колмогоров-Смирнов/ Лилиефорс намалява (при n = 500, D = 0.0269, p = 0.1500, а при n = 1 000, D = 0.01800, p = 0.1500) (Park, 2008, стр. 10). Това, както и цитираните по-горе изследвания на качествата на тестовете за нормалност показват, че тези тестове не се провалят, когато разпределението на променливата в генералната съвкупност е нормално. Следователно предположението, че получените в настоящото изследване резултати се дължат на една (негативна) особеност на използваните инструменти за проверка на нулевите хипотези за нормалност, е неоснователно.

От друга страна, макар и малко на брой, изследванията на реални данни показват, че разпределенията с нормална форма са по-скоро рядък прецедент, отколкото правило. Разпределенията на баловете от тестовете за постижения, а от други типове

психологически изследвания, се характеризират с най-различни степени на асиметричност, ексцес, тегла на опашките и модалности, което дава основание на Т. Мичери да заключи, че твърде малка част от разпределенията са „дори сравнително близка апроксимация на Гаусовото" (Micceri, 1989, стр. 161). (виж още Fan, 1998 и др.)

Далеч по-убедително обяснение на голямото разнообразие от форми на разпределенията, отличаващи се от тази на нормалното, може да се потърси в структурата на отделните извадки. Те са формирани ако не случайно (в статистическия смисъл на термина), то напълно стихийно. Във всяка една от тях са попаднали кандидати, които идват от различни краища на страната, принадлежащи към различни възрастови групи (не само току-що завършили средното си образование) и социални прослойки, с различна образователна история, с различна степен на общообразователна подготовка и мотивация за успех. С други думи, генералната съвкупност се характеризира с определена липса на хомогенност. В своя исторически преглед С. Стиглер подчертава, че статистиците след Ф. Галтън признават хомогенността на популацията като условие за формиране на нормално разпределение (Stigler, 1986). Всичко това способства за разкъсване на вътрешната, монолитна на пръв поглед, структура на извадките, които следва да се разглеждат като хетерогенни, съставени от различен брой субизвадки. Безспорно тази структура намира отражение в тестовите балове, следователно и в техните разпределения.

На трето място следва да обърнем внимание на структурата на способностите, чиито разпределения са обект на изследване. По-голяма част от разпределенията (общо 120, 90.91% от всички) са свързани с отделни раздели от ТОП (български език, литература, история и т. н.), които са предназначени за измерване на съответната способност. Останалите 12 (9.09%) представят общия тестов бал, интегриращ в себе си способностите от отделните раздели. Поради начина на неговото формиране, може да се предположи, че общият тестов бал представлява съвместно многомерно разпределение на отделните способности, което би могло да повлияе силно на формата на неговите разпределения От друга страна, очакването за субтестовите балове е да бъдат едномерни, отразяващи една единствена способност. Данните обаче сочат, че ако пренебрегнем резултатите от тестовете на съгласието, разпределенията на общия тестов бал не се характеризират със системно по-високи отклонения от Гаусовото разпределение в сравнение със субтестовите балове. Това означава, че бихме могли да очакваме многомерност и на субтестово равнище, или, като цяло, друга структура на способностите, различна от формално заложената в ТОП.

И накрая, отклоненията от нормалната крива вероятно се дължат на един вътрешноприсъщ „дефект" на тестовите балове. Тестовият бал е обобщена мярка на способностите, резултативна величина от използването на дадена сумарна скала, между айтемите на която съществува определена корелация. Парадоксът е в това, че корелацията между айтемите е желан ефект, върху който се базира надеждността на ска-

лата, но, както отбелязва Дж. Нунали, високите корелации от порядъка на 0.40 биха довели до разпределения, значително по-плоски от нормалното (Nunnally, 1978). Както беше показано по-горе, при разпределенията, анализирани в настоящото изследване, преобладават тези с отрицателен ексцес, който може да бъде обяснен с ефекта на взаимовръзките между отделните (суб)тестови въпроси. Т. Мичери добавя още един важен щрих – самите айтеми представляват отделни променливи, които имат свои кумулативни функции на разпределение и предположението, че всички те са Гаусови, изглежда неоснователно (Micceri, 1989).

Опит за генерализация

След като тестовете за нормалност на изследваните разпределения на резултатите от ТОП водят до отхвърляне на всички нулеви хипотези, при достатъчна подкрепа на тези решения и от други (числови и графични) методи, и при наличието на сравнително голям масив от разпределения (общо 132), върху които е базирано изследването, уместно ли е генерализирането на извода, че разпределенията на способностите се различават от нормалното? Или, ако се върнем към противоположните мнения на К. Хопкинс, Дж. Глас и Р. Гиъри, обича ли Бог нормалната крива или тя е сами мит?

Р. Тапиа и Дж. Томпсън поставят въпроса по противоположен начин – възможно ли е да се пренебрегнат такива резултати, получени от ограничен брой разпределения, с ограничен брой наблюдения, значително по-малък от обема на генералната съвкупност? (Tapia & Thompson, 1978)

Тук са възможни два алтернативни отговора. Първият от тях е, че тези резултати характеризират латентните променливи в ограничени по обем извадки и че това не означава непременно, че тяхното разпределение в генералните съвкупности не е нормално. Тази позиция се аргументира с тезата, че с нарастването на обема на извадката разпределението на тестовите балове се стреми към нормалното (Tapia & Thompson, 1978; Micceri, 1989). Авторите справедливо отбелязват, че тази теза отразява едно изопачено разбиране на централната гранична теорема. Съгласно този основен принцип при нарастване на обема на извадка, формирана от независими и случайни наблюдения, извадковото разпределението на сумата от стойностите (или на средните стойности) на множество такива извадки със същия обем, извлечени от същата генерална съвкупност, се стреми към нормалното разпределение. При това върху разпределението на наблюденията в генералната съвкупност не се налагат никакви ограничения, включително и по отношение на неговата форма, с изключение на условието на Линдберг за крайност на дисперсията (Калинов, 2010). Тази теорема обяснява кардиналното значение на нормалното разпределение при статистическия извод, но не дава основания за подмяна на средните стойности, чието разпределение се стреми към нормалност, с индивидуалните балове на и. л.

Алтернативният отговор на поставения въпрос, че ако разпределението на ге-

нералната съвкупност не е нормално, е малко вероятно (простите случайни) извадки да придобият нормална форма. Като основно свидетелство можем да посочим медианата на Тюки, съгласно която типичното извадково разпределение не е нормално. Това предположение е по-правдоподобно, защото не противоречи на централната гранична теорема. То се съгласува с огромна част от резултатите от приложените в настоящото изследване методи за оценка на нормалността, както и на данните от литературните източници. Беше показано обаче, че в някои случаи, главно чрез графичните методи, но и при прилагането на консервативните тестове на съгласието, някои разпределения могат да бъдат оценени като нормални. Следователно имаме основания да допуснем, че разпределението в генералната съвкупност е близко до нормалното, но се отклонява от него.

Ако нормалното разпределение е мит, то значи ли това, че Бог го не го обича? Струва ни се,че след всичко казано дотук можем да заключим в същия афористичен стил, че Бог може би е постановил принципа на нормалността, но не се интересува от неговото спазване.

Как трябва да се постъпи в случаите, в които разпределенията на тестовия бал се отклонява от нормалното?

Когато променливата не е нормално разпределена, е необходимо нейните стойности да бъдат трансформирани така, че новото разпределение да бъде нормално. При нормализиране на разпределението се променя скалата на измерване, което води до формиране на нова променлива, математически еквивалентна на изходната, но с нормално разпределение. Основните методи за нормализиране на разпределенията са чрез използване на втори корен от стойностите или (за крайно асиметрични разпределения) логаритмична трансформация (с основа натуралния логаритъм e или с основа 10). Прилага се също и инверсна трансформация $\left(x_i^{-1}\right)$(Gardner, 1975; Weiner et al., 2003).

Много изследователи обаче предупреждават, че използването на нормализирани разпределения трябва да бъде внимателно поради необходимостта от правилен избор на метод на трансформация и проблеми с интерпретирането на резултатите (Taylor, 1985; Micceri, 1989).

Изследване 2. Анализ на латентните структури на Теста по общообразователна подготовка

2.1. Цели на изследването

Едно от основните изисквания за прилагане на по-голяма част от моделите на IRT, както и на моделите на CTT, включително и на базовите модели, е едномерността на теста, което означава, всички въпроси в него трябва да измерват различни аспекти на една и съща способност. Ако това е така, въпросите могат да бъдат представени като точки, наредени на един континуум, представящ съответната способност. Тук е уместно да припомним двата основни постулата на Л. Л. Търстоун за психологическите измервания: (1) „..измерването описва само един атрибут на обекта" и (2) „…друг постулат, който лежи в основата на всички измервания е, че измервания атрибут е винаги едномерен" (Thurstone, 1929, стр. 158-159). Определянето на латентната структура на теста е „критичен въпрос" при моделирането на тестовите данни, независимо от това дали се обсъжда прилагането на едномерен или многомерен модел (Embretson & Reise, 2000, стр. 227; de Ayala, 2009). Под „латентна структура" на теста ще разбираме съвкупността от скрити, ненаблюдаеми пряко променливи (димeнсии), за които може обосновано да се предположи, че оказват влияние върху отговорите на и. л. по отделните въпроси от теста.

Прилагането на неадекватни тестови модели, особено на такива, които не са съобразени с размерността на пространството на латентните черти, е начинание, свързано с риска от допускане на грешки. Ако латентната структура е многомерна „по природа", с прилагането на едномерен модел може да бъде нарушено едно от основните допускания на IRT – допускането за локална независимост. Това би довело до изместване на оценките на параметрите на въпросите, от една страна, и до увеличаване на стойностите на стандартните грешки, свързани с оценките на различните нива на способност, от друга (Weiner et al., 2003). Следователно преценката за размерността на латентната структура, лежаща в основата на теста, не може да бъде направено *a priori*, дори и въз основа на неговото предназначение или обичайна употреба. Поради това е необходимо да се направи предварително изследване на латентната структура на ТОП, при допускане, съгласно базовите модели, за нейната едномерност.

Проучването се провежда, за да се потърсят доказателства „за" или „против" прилагането на едномерен модел към данните от ТОП, и по-конкретно - да се потърсят отговорите на следните изследователски въпроси:

(1) Каква е размерността на латентното пространство на способностите, които се измерват чрез ТОП (чрез колко латентни променливи може да бъде обяснена мре-

жата от взаимовръзки между компонентите на теста)?

(2) Колко добре тези латентни фактори обясняват тестовите резултати?

(3) Ако не се наблюдават едномерни латентни пространства в строгия смисъл на понятието, има ли главни и второстепенни фактори и има ли отчетливо разграничение между тях?

(4) Ако (главните) фактори са повече, има ли сред тях един, който да бъде идентифициран като доминиращ и който да обоснове допускането за едномерност на латентните структури на теста?

(5) Каква е същността на тези латентни фактори?

2.2. Методология

2.2.1. Дизайн

Тъй като липсват резултати от предходни експериментални изследвания, които да послужат като основа за предварително определяне на броя и структурата на латентните променливи на ТОП, ще разглеждаме допускането за нейната едномерност като работно, а не като хипотеза в строгия смисъл на термина.

Изследването на латентната структура на ТОП, а и на всеки психометричен инструмент за измерване, по същество е каузално, предполагащо установяване на взаимовръзките между определен брой независими и зависими променливи. Нещо повече, то е фокусирано върху същината на независимите променливи. Поради това, че те не са пряко измерими, изследването по необходимост приема формата на анализ на мрежата от взаимовръзки между зависимите променливи.

В качеството на зависими променливи величини ще разглеждаме отделните манифестирани променливи, т. е. постиженията (тестовите резултати), постигнати от изпитаните лица по различните предметни области, регистрирани чрез съответните компоненти на теста. По-конкретно, като зависими променливи ще се разглеждат отговорите на изпитаните лица на отделните тестови въпроси на две равнища: (1) в рамките на всеки субтест, т.е. броят на компонентите във всеки анализ е 10, и (2) на равнище цялостен тест, т.е. броят на компонентите във всеки анализ е 100.

Приемаме, че наблюдаваните променливи на равнище тестов въпрос са дихотомизирани манифестации на континуалните, метрични ненаблюдаеми способности на и. л., които са обект на анализ. За да се разкрие латентната факторна структура, е необходимо да се определят корелациите между наблюдаваните променливи. Поради това, че като компоненти на теста се разглеждат отделните въпроси, ще бъде използван тетрахоричен коефициент на корелация (предназначен за анализ на дихотомизирани променливи) вместо r на Пиърсън (приложим за интервални скали) или ϕ (за дискретни бинарни скали). В сравнително изследване на различни техники за факторизиране на бинарни данни, Д. Кнол и М. Бергер правят заключението, че тетрахоричният

коефициент работи толкова добре, колко и останалите методи на корелация (Kim & Mueller, 1978; Knol & Berger 1991).

В качеството на независими променливи величини ще разглеждаме латентните променливи, за които може да се предположи, че въздействат върху манифестирани променливи. В този смисъл те могат да бъдат разглеждани и като предиктори на тестовите резултати. Изходната позиция, от която тръгваме, е, че броят на независимите латентни променливи е неизвестен, по-конкретно, че всяка зависима променлива би могла да бъде свързана с която и да е от латентните променливи. Поради това дизайнът на изследването може да бъде представен като търсене на връзката между m латентни независими променливи (фактори) F и n наблюдавани зависими променливи X. Броят на латентните променливи m може да варира в интервала $1 \leq m \leq n$.

2.2.2. Данни

Изследванията върху латентната структура на ТОП бяха извършени последователно върху набор от 12 варианта на теста, като за основна структурна единица на всеки анализ бе определен (1) отделният субтест с обем от 10 въпроса и (2) целия тест с обем от 100 въпроса. Като сурови данни са използвани дихотомично кодираните отговори на и. л. на всеки въпрос (1 при коректен отговор и 0 при всички останали случаи).

Основанията за това са следните. От една страна, ТОП е разработен, както бе отбелязано по-горе, като инструмент за оценяване, предназначен за проверка и оценка на способностите на кандидат-студентите в 10 различни предметни области. В основата на тази концепция стои идеята, че всеки субтест на ТОП измерва една единствена способност в съответната предметна област (езикова, литературна, математическа и т. н.), която е самостоятелна и независима от способностите в другите предметни области на теста. Поради това резултатите от отделните субтестове се използват като самостоятелни балообразуващи компоненти при класирането на кандидатите.

От друга страна, общите резултати от теста също се използват като балообразуващ компонент. Правдоподобно е да се предположи, че скритата структура на теста не съвпада с очевидната му предметна структура, а да е организирана на друг принцип. Могат да се предположат, например, взаимовръзки между резултатите по български език и литература, по история и география, по математика и физика или между тези по химия и биология, които в средното образование се разглеждат и изучават като културно-образователни области, в рамките на всяка от които функционират единни Държавни образователни изисквания за учебно съдържание. В този смисъл, всяка съвкупност от въпроси може да се разглежда като случайна съвкупност, манифестираща определен брой области. Въпросът, който стои пред психометрика, като отбелязва В. Ревел, е да определи колко такива области са представени в извадката и колко добре всеки въпрос представя тези области (Revelle, 2011).

Съобразно структурата на ТОП, подходящо би било анализът на данните да бъде извършен на две равнища: (1) от гледна точка на ТОП като тестова батерия да се изследва факторната структура на отделните субтестове и (2) да се проучи факторната структура на ТОП като цялостен тест.

2.2.3. Методи за анализ на данните

В търсене на размерността на латентната структура на ТОП могат да се използват различни подходи, методи и свързаните с тях индикатори. С. Ембретсън и С. Рийз посочват над 10 метода за оценка на размерността (Embretson & Reise, 2000). Р. Нандакумар също представя доста подробен преглед на изследователските методи, предназначени за тази цел (Nandakumar, 1993). Голяма част от тях се базират на идеята за търсене на някакъв тип взаимовръзка между компонентите[7] на теста, т.е. между манифестираните променливи, която да бъде обяснена чрез влиянието на една или повече латентни променливи.

Един от тези подходи е да се направи оценка на вътрешната консистентност на теста, която се разглежда като хомогенност на неговите компоненти; като степен, в която тези компоненти са ориентирани към един и същи конструкт (латентна черта или характеристика на индивидите). Мнозина автори дефинират вътрешната консистентност именно в този план - като степен, в която "…всички въпроси измерват (т. е. са манифестация на) едно и също нещо" (DeVellis, 2003, стр. 28).

Вътрешната консистентност на теста е характеристика, която се основава на корелацията между отделните му компоненти. Ако отговорите на и.л. на въпросите в един тест, предназначени за измерване на един и същи теоретичен конструкт, са съгласувани, т.е. корелират силно, този тест има висока вътрешна консистентност. В областта на психологическите измервания са разработени няколко метода за оценка на надеждността на тестовите резултати, основани на вътрешната консистентност на въпросите в теста. Сред най-често използваните мерки, поради своите безспорни качества, е коефициентът α на Кронбах. Стойностите на този индекс са функционално зависими както от броя на компонентите в теста, така и от (средната) корелация (*mean inter-item correlation*) между тях.

Тези особености на α дават основание на мнозина автори да приемат, че коефициентът дава оценка на това колко добре едно множество от въпроси (променливи) измерва един-единствен, едномерен латентен конструкт, и да го използват като мярка за едномерността на данните (Miller, 1995; Wiberg, 2004). Логиката, на която се гради това схващане е, че стойността на α нараства тогава, когато расте корелацията между отделните въпроси. Високата корелация е белег за това, че въпросите са ориентирани

[7] Като компоненти на измервателния инструмент могат да се разглеждат както отделните въпроси, така и неговите по-големи структурни единици – групи от еднородни въпроси или субтестове (субскали).

към един и същи конструкт. Следователно, ако се наблюдават високи стойности на коефициента, това би било доказателство, че латентната структура е едномерна.

Разгледаните по-горе особености на консистентността и на един от подходите за нейната оценка - α на Кронбах, изглежда правят тази статистика подходящ кандидат за определяне на размерността на ТОП. Има обаче сериозни теоретични възражения срещу използването на α като тест за едномерност. Едномерността на дълбинната структура е необходима предпоставка за постигане на точна оценка на вътрешната консистентност, и обратно – не е извод, който може да бъде направен въз основа на наблюдаваните високи стойности на коефициента (Miller, 1995; Graham, 2006)

Коефициентът α на Кронбах е свързан концептуално и емпирично с друг популярен статистически метод за експлициране на латентни променливи - факторния анализ. Родствената връзка между е тях е в това, че и двата метода експлоатират корелационните връзки между компонентите на психометричния инструмент. Но между двата метода има и една принципна разлика, която дава огромно предимство на факторния анализ като средство за търсене на размерността на латентната структура в данните: въпреки че едномерността на теста е необходимо условие за прилагане на α като неизместена оценка на вътрешната му консистентност (надеждност), равнището на коефициента не е свързано с факторната структура (хомогенност) на теста. Аргументът за това е, че стойността на α зависи от големината на средната корелация $\bar{r}$, докато размерността на латентното пространство се определя от структурата на корелационните отношения между въпросите.

Като алтернативи на факторния анализ могат да се разглеждат и такива многомерни статистически техники като многомерното скалиране и клъстърния анализ, а защо не и анализа на съответствията, които също се използват за редуциране на данните. Още повече, че тези методи боравят с много по-широк спектър от различни видове матрици на сходства и различия (*similarity/ dissimilarity matrices*). Корелационните отношения също могат да се разглеждат като тип сходство и такива матрици също се подлагат на споменатите по-горе многомерни анализи. Но, както отбелязва Р. Дарлингтън, корелациите имат някои свойства, към които другите многомерни техники не са чувствителни (Darlington, 1997). Едно от тях е реверсивността на зависимите променливи. Ако дадена променлива бъде „обърната" (т. е. стойностите й бъдат заменени с техните противоположни по смисъла на съответната скала), корелационните коефициенти на тази променлива с останалите няма да променят големината, а само знака си. Това може да промени и знака пред факторното тегло на съответната променлива, което също няма особено значение при неговата интерпретация. Различният знак пред конкретна променлива за конкретен фактор има специфичен смисъл – той показва само, че тази променлива е свързана с фактора по противоположен начин (Kim & Mueller, 1978). Такава промяна би променила съществено резултатите от многомерното скалиране или клъстърния, но не и при факторния анализ. Многомерното скалиране, отбе-

лязва Р. Дарлингтън, разкрива фактори, които диференцират променливите, докато факторният анализ търси фактори, които лежат в тяхната основа (Darlington, 1997). Поради дискутираните по-горе негови особености факторният анализ бе предпочетен като методология за изследване на латентната структура на ТОП.

2.2.3.1. Избор на вид факторен анализ

Наименованието „факторен анализ" обхваща широка група от статистически многомерни аналитични техники, които имат две основни сфери на приложение: (а) за намаляване на броя на наблюдаваните променливи, т.е. редуциране на наличните данни (*data reduction method*) и (б) за разкриване на структурата на взаимовръзките между променливите, т.е. за класифициране на променливите и конструиране на скали (*structure detection method*). Методът се заражда в недрата на психометрията във връзка с изследванията на интелигентността, като особен принос за развитието и утвърждаването му има Л. Л. Търстоун (Thurstone, 1931b; 1934; 1935; 1947), но подобни техники са използвани за първи път от Ч. Спиърмън още в началото на миналия век във връзка с класическото му изследване на интелигентността, в което той изследва резултатите от различни тестове за умствени способности. В като обобщение на получените резултати той предполага, че огромното разнообразие от умствени способности – математически, вербални, художествени и логически, могат да бъдат обобщени с един латентен фактор на обща интелигентност g (Spearman, 1904), който се проявява заедно с множество други специфични фактори.

В зависимост от предпоставките, на които се базира факторният анализ, както и от неговите цели, се разграничават два основни типа: изследователски факторен анализ (*exploratory factor analysis, EFA*) и потвърдителен факторен анализ (*confirmatory factor analysis, CFA*). Изследователският факторен анализ е подходящ за случаите, в които липсва ясна теоретична база или резултати от предходни експериментални изследвания относно природата на латентната структура, поради което неговата цел е изграждане на факторен модел, който да обясни взаимоотношенията между наблюдаваните променливи (Revelle, 2011). Поради това *EFA* се разглежда предимно като процедура за генериране на теоретични модели и като анализ, подходящ за разработване на психологически скали (Kubinger, 2003). Обратно, потвърдителния факторен анализ се провежда въз основа на ясни теоретични или емпирични основи, а неговата цел е верифицирането на конкретен факторен модел. Поради това *CFA* се разглежда предимно като процедура за тестване на разработени вече факторни модели, т. е. на специфични хипотези за факторната структура на група от променливи (Hurley et al., 1997; Stevens, 2002).

Поради обстоятелството, че настоящото изследване на факторната структура на ТОП е първото по рода си и се провежда в условия, които съответстват на онези, подходящи за прилагането на изследователския тип факторен анализ, при анализа на

данните първоначално ще пристъпим към изграждане на подходящ модел (или подходящи модели) на латентната структура на ТОП, който впоследствие ще бъде верифициран чрез потвърдителен факторен анализ.

2.2.3.2. Избор на метод за факторизиране

В съответствие с очертаните по-горе предназначения на метода изследователският факторен анализ съществува в две основни форми: анализ на главни компоненти (*Principal component analysis*, *PCA*) и анализ на главни фактори (*Principal factor analysis*, *PFA*). Двата подхода споделят обща философия за факторизиране на наблюдаваните променливи - разбиването им на подмножества от променливи, свързани помежду си с висока корелация, и „обединяването" на променливите във всяко подмножество в обща дълбинна променлива (фактор). Крайната цел на факторния анализ може да бъде определена като постигане проста факторна структура (*simple structure*), която се поддава лесно на интерпретация (Thurstone, 1935, 1947; Kim & Mueller, 1978; Reynolds & Kamphaus, 2003). Двата метода са подходящи за приложение в ситуации, в които размерността на данните и тяхната структура не са добре познати. Поради това *PCA* и *PFA* се отнасят към групата на многомерните изследователски техники (*Multivariate exporatory techniques*), които са сред най-често използваните в рамките на *EDA*, т.е. не се базират на формулиране на статистически хипотези (например относно броя на факторите) и тяхната верификация (McDonald, 1985). Най-често прилаган е линейният факторен анализ, макар и да са разработени и нелинейни аналитични модели (de Ayala, 2009).

Макар и да преследват една и съща цел, между двата подхода за факторизиране има и някои съществени различия, познаването на които би подпомогнало избора на подходящия метод. Анализът на главни компоненти е предназначен за комбиниране на наблюдаваните променливи в малък брой подмножества (главни компоненти), при което се отчита цялата дисперсия (*total variance*) на отделните променливи. Поради това всеки извлечен компонент представлява линейна комбинация на изходните променливи, която е по-скоро геометрична абстракция, която не винаги съответства на определена психологическа променлива. Тези компоненти, според някои автори, имат нищожно теоретично значение (Hakstian & Muller, 1973).

Анализът на главни фактори се основава на доста по-различен методологичен подход. Той се базира на разбирането, че отговорите на и. л. по всяка отделна променлива са зависими от (съответно дисперсията на тази променлива отразява) три компонента. Първият от тях е един (или повече) общи латентни фактори (*common factors*), които оказват влияние върху всички променливи. Такъв общ фактор би могъл да бъде например „разбиране при четене", т.е. разбиране и извличане на смисъла на тестовия въпрос, който обикновено представлява кратък писмен текст. Отговорът на изпитваното лице на всеки тестов въпрос, независимо към коя предметна област при-

надлежи, е зависим от това доколко добре, доколко правилно това лице е разбрало поставения въпрос. Така векторът на отговорите на това, а и на всички и. л., търпи влиянието на общите фактори и част от неговата дисперсия (*common, shared variance*) може да бъде обяснена с това влияние. Вторият компонент е някакъв специфичен, уникален аспект от съответната предметна област, който характеризира даден конкретен въпрос и нито един от останалите. Пример за такова единствено по рода си знание може да бъде въпросът за първата владетелска династия, управлявала България до средата на VIII век. Третият компонент отразява грешката на измерването.

Разграничителната линия между двата основни метода на факторизиране минава между компонентите на дисперсията на отделните въпроси. Докато при анализа на главни компоненти се третира, както бе отбелязано, цялата (*total*) дисперсия на променливите, при анализа на главни фактори се използва само онази нейна част, която се дължи на общите фактори и която те споделят помежду си (*communalities*). Частта от дисперсията, която е уникална за дадена променлива, се дефинира като разлика между цялата дисперсия, от една страна, и споделената дисперсия и дисперсията на грешката – от друга. Поради това като първоначална оценка на споделената дисперсия на даден въпрос при анализа на главни фактори обикновено се използва квадратът на коефициента на множествена корелация (*multiple* R^2_j) от регресията на *j*-тия въпрос с всички останали въпроси в измервателния инструмент (McDonald, 1985; Buja & Eyuboglu, 1992).

Поради тези особености на двата метода, при решаването на изследователски задачи, свързани с редуциране на данните (броя на променливите), предимство има анализът на главни компоненти, а за разкриване на латентни структури – анализът на главни фактори. Тъй като при втория метод се анализира общата, споделена от променливите дисперсия, извлечените фактори могат да се концептуализират и идентифицират като психологически променливи.

Задачата за определяне на факторната структура на ТОП предполага не просто редуциране на броя на неговите компоненти (въпроси или субтестове), а разкриване на латентната му структура и нейната съдържателна интерпретация. Независимо от ясното разграничение между двата метода на факторизиране, изследователите не са единодушни за сферите на тяхното приложение. Някои споделят мнението, че към едни и същи данни могат да бъдат приложени успешно различни факторни аналитични техники (Yates, 1987; Velicer, Eaton & Fava, 2000; de Ayala, 2009), други възприемат анализа на главни компоненти като ”неистински метод за факторен анализ” (Costello & Osborne, 2005; Hayton, Allen & Scarpello, 2004). Независимо от тези разногласия, въз основа на характеристиките на двата метода, представени по-горе, като метод за факторизиране на променливите в настоящото изследване бе предпочетен анализът на главни фактори.

Общият модел на този вид факторен анализ може да бъде представен по след-

ния начин (по Harman, 1976, стр. 15):

$$z_j = a_{j1}F_1 + a_{j2}F_2 + \ldots + a_{jm}F_m + u_jY_j \quad (60)$$

където:

z_j – j-та наблюдавана променлива, $j = 1, 2, \ldots, n$

F – общи фактори ($m \leq n$)

Y – специфични (уникални) фактори

a – факторни коефициенти (тегла)

В този модел всяка манифестирана променлива е представена като линейна функция от няколко общи фактора, обясняващи споделената корелация между манифестираните променливи, а всеки специфичен фактор обяснява останалата дисперсия, включително и грешката на измерване.

В рамките на анализа на главни фактори са разработени различни аналитични техники за оценка на параметрите на модела – Communalities, MinRes (*Minimum Residual factor method*), Maximum likelihood factors, Centroid method, Principal axis method и др. Мнозина изследователи разглеждат тези техники като сходни и същевременно алтернативни методи за постигане на едни и същи цели (Harman & Jones, 1966; Harman, 1976; Reynolds & Kamphaus, 2003). Трудно е обаче, според А. Йейтс, да се прецени е какво би било отражението на избора на една или друга техника върху изводите, които ще бъдат направени от съответния анализ. Авторът отбелязва, че всяка промяна в математическите критерии при прилагането на изследователския факторен анализ би имала съществено отражение върху резултатите (Yates, 1987).

Това мнение намира различни потвърждения, например в сравнителното изследване, направено от Р. Торндайк, на годността на няколко метода за извличане на фактори да водят до постигане на проста факторна структура. Авторът съпоставя 5 метода, приложени върху 24 корелационни матрици, извлечени от 6 научни области извън психологията (6 области Х 4 корелационни матрици). Изводът на автора е, че методите за факторизиране се различават по своя потенциал за продуциране на простата факторна структура при различни данни, но като цяло MinRes и Maximum likelihood factors дават по-добри решения (Thorndike, 1971, стр. 4).

Други автори също отдават своите предпочитания на метода на максималното правдоподобие, който е подходящ за приложение в случаите, в които разпределенията на променливите са нормални или не се отличават съществено от нормалното. В останалите случаи авторите препоръчват като по-подходящ метода на главните оси (*Principal axis factoring*), при който няма изискване за формата на разпределенията (Fabrigar, Wegener, MacCallum & Strahan, 1999; Costello & Osborne, 2005; Revelle, 2007).

След внимателно проучване на особеностите на различните техники за факторизиране бе направен избор в полза на метода на главните оси, който се основава на

следните аргументи:

(1) Различните аналитични техники като цяло водят до сравними резултати, тъй като анализът на главни фактори като генеричен метод за извличане на факторите е сравнително устойчив срещу влиянието, което прилагането на една или друга техника може да окаже върху резултатите (Zwick & Velicer, 1986).

(2) Резултатите от собствените пилотни изследвания върху субтестовите данни, направени с различни факторни техники в рамките на PFA, потвърдиха мнението на У. Цвик и У. Велисер. Макар и в някои случаи с различен брой фактори или с различни собствени стойности, извлечените по различен начин факторни конфигурации на едни и същи субтестове са напълно съпоставими.

(3) Прилагането на този метод се основава на допусканията, общи за факторния анализ, но не и за формата на разпределенията на променливите. Поради представените в различни литературни източници доказателства, че разпределенията на променливите в социалните и поведенческите науки се отклоняват, понякога съществено, от нормалното разпределение, този метод е по-подходящ за данните от ТОП от, например, метода на максималното правдоподобие.

(4) Необходимо е съгласуване на метода за извличане на факторите и метода за определяне на оптималния им брой. Поради използването на специфичния метод на паралелния анализ, описан в следващата глава, най-подходящ за извличане на факторите е методът на главните оси.

2.2.3.3. Избор на метод за определяне на броя на факторите

Една от основните задачи, свързани с използването на изследователския факторния анализ, е решаването на т. нар. „проблем за броя на факторите" (*number-of-factors problem*). Той се състои в определяне на „оптималния" брой на факторите от гледна точка на главната цел на факторния анализ – редуциране на броя на изходните променливи и създаване на модел с малък брой, ясни и поддаващи се на интерпретация фактори и в същото време съхраняване на важната информация чрез пълноценно представяне на мрежата от корелационни връзки между променливите. Това е може би най-важното, но и най-трудно изследователско решение, тъй като, както отбелязват К. Рейнолдс и Р. Камфаус, "няма твърдо установени правила и поради това процесът [на вземане на решение] е смесица от навици, субективност и преценка" (Reynolds & Kamphaus, 2003, стр. 85). На този проблем са посветени множество публикации (виж Thorndike, 1971; Yates, 1987; Fabrigar et al., 1999; Kabacoff, 2003; Hayton, Allen & Scarpello, 2004; Ledesma & Valero-Mora, 2007 и др.). От „техническа" гледна точка, това решение е по-важно от избора на метод за извличане на факторите или за тяхната ротация, срещу влиянието на които *PFA* е сравнително устойчив (Zwick & Velicer, 1986). От съдържателна гледна точка този процес е важен при установяване на конструктната валидност на измервателния инструмент, която се установява чрез определяне на

пространството на латентните променливи (Nunnally, 1978). Съществуват редица емпирични доказателства, че както надценяването, така и подценяването на броя на факторите влияе негативно върху оценката на факторните тегла, а следователно и върху интерпретацията на факторите (Velicer, Eaton & Fava, 2000, по Hayton, Allen & Scarpello, 2004).

Поради важността на проблема, за неговото решаване са разработени множество подходи, които могат да бъдат използвани при вземането на решение. Трябва да се отбележи обаче, че между изследователите няма съгласие кои методи са най-подходящи, а и паралелното им прилагане често пъти води до противоречиви резултати (Zwick & Velicer, 1986; Hayton, Allen & Scarpello, 2004; Revelle, 2007).

Може би най-популярният сред тях е критерият на Х. Кайзер (*Kaiser-Guttman rule „greater than one"*) $K1$, съгласно който във факторния модел следва да бъдат включени само онези фактори, чиято собствена стойност (*eigenvalue*) надхвърля 1.00 (Kaiser, 1960; Fabrigar et al., 1999; Reynolds & Kamphaus, 2003; Hayton, Allen & Scarpello, 2004). Смисълът на това изискване е, че това е стойността на цялата дисперсия на една отделна манифестирана променлива във факторния анализ. Следователно, ако даден фактор обяснява по-малко дисперсия, отколкото съдържа една променлива, той следва да бъде изключен от модела. Х. Кайзер осмисля това правило въз основа на доказателството на Л. Гутман (Guttman, 1954), че най-ниският ранг на една популационна корелационна матрица (равен на минималния брой компоненти, които обясняват корелациите извън главния диагонал) може да се получи чрез извличане само на факторите със собствена стойност по-голяма от 1.00 (Hayton, Allen & Scarpello, 2004). Солидната теоретична основа, както и лесната приложимост на $K1$ обясняват неговата повсеместна употреба, включително и като основен критерий в редица статистически софтуерни пакети.

Друг популярен метод е графичният тест на Р. Кетел (*Cattell's scree test*). Тестът представлява двумерна графика (начупена линия) на собствените стойности (разположени по ординатата), срещу номерата на последователно извлечените фактори (представени на абсцисата). Авторът използва образно един геоложки термин (*scree* – "сипей"), тъй като обикновено начупената линия има формата на това геоложко образувание – тя е силно скосена в лявата част на графиката, след което рязко приема похоризонтална форма в дясната й част. Базирайки се на метода *Monte Carlo*, Р. Кетел препоръчва броят на факторите, които следва да се оставят в моделираната латентна структура, да се определи според мястото на точката, в което стръмно спускащата се линия на собствените стойности рязко променя посоката си и преминава в слабо наклонена, почти хоризонтална начупена линия. Вдясно от тази точка на пречупване, според Р. Кетел, се намират „факторните отломки" – това са факторите, които са уловили случайния „шум" в данните и които следва да бъдат изключени от модела (Cattell, 1966; Kim & Mueller, 1978; Reynolds & Kamphaus, 2003; Hayton, Allen & Scarpello, 2004;

Ledesma & Valero-Mora, 2007).

Според някои автори слабостите на двата метода са повече от техните предимства, а тяхната популярност се дължи на обстоятелството, че именно те са включени в ресурсите на повечето статистически програми (Kabacoff, 2003; Costello & Osborne, 2005; Ledesma & Valero-Mora, 2007).

Като основна слабост на критерия на Кайзер-Гутман неговите многобройни критици отбелязват,че при прилагането му се забелязва ясна тенденция за значително надценяване на броя на факторите. Това „поведение" на критерия се дължи на обстоятелството, че правилото на Л. Гутман се отнася за популационните матрици (Horn, 1965). Но дори и всички собствени стойности на една популационна корелационна матрица да са равни на 1.00 и нито една да не надвишава тази стойност (например, ако променливите в популацията са независими), всяка ограничена по обем емпирична извадка може да генерира, поради извадковата вариативност, собствени стойности, по-големи от 1.00 (Buja & Eyuboglu, 1992). Следователно, при извадковите корелационни матрици извадковата грешка се добавя към ранга на матрицата, увеличавайки броя на факторите. Като друга слабост на критерия на Кайзер-Гутман се посочва, че той е механично правило, което може да доведе до произволни решения, например фактор със собствена стойност 1.01 е значим, но не и фактор със собствена стойност 0.99 (Fabrigar et al., 1999). Забелязва се системна връзка между броя на анализираните променливи и броя на факторите, който е между 1/3 и 1/5 до 1/6 от този на променливите (Ledesma & Valero-Mora, 2007). Най-важната особеност на метода на Кайзер-Гутман е, че той е предназначен за оценка на броя на значимите фактори в контекста на Анализа на главни компоненти, при който се отчита цялата дисперсия на променливите, с единици в главния диагонал на корелационната матрица.

Графичният метод на Р. Кетел е критикуван заради субективността при определяне на точката на пречупване на графиката, особено в случаите, в които могат да бъдат наблюдавани няколко такива точки (Reynolds & Kamphaus, 2003). От друга страна, на изследователите с практически опит нерядко се налага да анализират графики, в които собствените стойности намаляват с (почти) равна стъпка и поради това такива точки на пречупване въобще не могат да бъдат идентифицирани

Разработени са редица методи, алтернативни на горните два, но голяма част от тях почиват на тяхната идеология - базирани са или на големината на собствените стойности, или на съотношенията (дистанциите) между тях. Така например Ф. Лорд предлага два критерия за определяне на едномерността чрез съпоставяне на латентните фактори: (1) ако първия фактор е „голям в сравнения с втория" и (2) ако вторият фактор не е „много по-голям от който и да е от останалите", тогава въпросите в теста са приблизително едномерни (Lord, 1980, стр. 21). Ф. Лорд отнася тези критерии към собствените стойности, получени от матрици на тетрахоричните корелации, и не фиксира конкретно нито техните големини, нито съотношенията между тях. В свое емпи-

рично изследване Р. Де Аяла прилага тези критерии, макар и върху резултати от матрица на Пиърсънови корелации (по-точно на φ), по подобен начин: въз основа на съждението, че разликата между първата и останалите собствени стойности е сравнително голяма (de Ayala, 2009). Други автори конкретизират предложенията на Ф. Лорд, представяйки ги като общи критерии за приемлива едномерност по следния начин: ако първия фактор обяснява 20% или повече от дисперсията или ако съотношението между собствените стойности на първия и втория фактор е 3:1 или дори 4:1 (Reckase, 1979; Cooke et al., 1999; Pollard et al., 2009).

Алтернативни методи за определяне на факторната структура са разработеният от И. Джолифе критерий (*Jolliffe's criterion*), който, приемайки, че правилото на Х. Кайзер води до запазване на твърде малко компоненти, смекчава строгите изисквания на *K1*, свеждайки долната граница на собствените стойности на компонентите до 0.70 (Jolliffe, 2002). В свое изследване Е. Холт и сътрудници използват като прагов критерий, наред с тестовете на Х. Кайзер и Р. Кетел, средната стойност на наблюдаваните собствени стойности (Holt et al., 2008, по Riekert & Eakin, 2008). В. Велисер развива теста *MAP* (*Minimum average partial*), също предназначен за приложение при Анализа на главни компоненти (Velicer, 1976; Hayton, Allen & Scarpello, 2004). Тестът на М. Бартлет (*Bartlett's test*) се основава на индекса на съгласието (*goodness-of-fit*) χ^2. М. Бартлет приема, че извличането на факторите трябва да продължи дотогава, докато корелационните коефициенти в матрицата на остатъчните корелации не достигнат случайно равнище. В този случай може да се приложи тестът χ^2, при което стойностите на критерия χ^2 заедно с асоциираните равнища на значимост (*p-levels*) се определят като функция от броя на извлечените фактори. Както показва изследването на В. Ревел, тази процедура води до извличането на твърде голям брой фактори (Bartlett, 1950; Hayton, Allen & Scarpello, 2004; Revelle, 2007). Друг подход за определяне на факторната структура, разработен от В. Ревел и Т. Роклин, е критерият за определяне на „най-простата структура" (*very simple structure*, *VSS*). Той формализира утвърдената практика, според която при интерпретиране на факторите изследователите се фокусират върху променливите с най-високо факторно тегло и пренебрегват онези с ниско факторно тегло. Основава се на съпоставяне на емпиричната корелационна матрица с тази, която може да бъде възпроизведена от една или друга опростена версия (S_{ck}) на изходната факторна матрица (F), получена от емпиричните данни (Revelle & Rocklin, 1979).

Специално внимание заслужава Паралелният анализ (*Parallel analysis*, *PA*), разработен от Дж. Хорн (Horn, 1965), който също реферира към двата основни метода за определяне на броя на факторите и е подходящ за приложение както при *PCA*, така и при *PFA* (Buja & Eyuboglu, 1992; Reynolds & Kamphaus, 2003). Дж. Хорн предлага да се генерират множество „изкуствени" съвкупности от данни (корелационни матрици), паралелни на емпиричната от гледна точка на броя на променливите и обема на извад-

ката, които да се моделират по случаен начин. За симулиране на данните се използва методът *Monte Carlo*, като променливите се извличат от нормално разпределени генерални съвкупности при условие (нулева хипотеза) за липса на статистическа зависимост между тях. При това условие всички собствени стойности в генералната съвкупност при Анализа на главните компоненти ще бъдат равни на 1.00, а при Анализа на главните фактори – 0.00 (Buja & Eyuboglu, 1992).

Множеството нови, симулирани съвкупности от данни се подлагат, успоредно със съществуващата емпирична съвкупност, на факторен анализ от същия тип и от всяка съвкупност се извличат поредните фактори с техните собствени стойности. В оригиналната версия на Дж. Хорн критерият за определяне на оптималния брой на факторите в емпиричните данни се основава на съпоставянето на собствената стойност на даден емпиричен фактор (например първия по ред) със средната стойност на собствените стойности на съответните симулирани фактори (всички първи по ред фактори от генерираните по случаен начин съвкупности). Правилото, предложено от Дж. Хорн, е в модела да се включат само онези емпирични фактори, чиито наблюдавани собствени стойности надвишават средната на „очакваните" собствени стойности на съответния пореден фактор, получени на случайно равнище (Buja & Eyuboglu, 1992; Glorfeld, 1995; Kabacoff, 2003; Hayton, Allen & Scarpello, 2004; Ledesma & Valero-Mora, 2007). Правилото на Дж. Хорн означава, че в оригиналния си вид неговият метод работи на равнище на значимост 0.50, което, както ще бъде показано по-нататък, го прави твърде либерален и води до надценяване на броя на факторите.

Паралелният анализ, както става ясно от неговата процедура, може да се разглежда като развитие на метода на Кайзер-Гутман. Той се основава на идеята, че собствените стойности на латентните променливи, извлечени от емпирични данни, могат да бъдат по-високи от тези, извлечени от данни, генерирани на случайно равнище. Поради това той е съществен опит да се преодолее вроденият недостатък на критерия на Кайзер-Гутман: базиран на правило, отнасящо се за популационни корелационни матрици, той системно надценява броя на факторите поради натрупването на извадковата грешка, поради което е подходящ за извадки, клонящи към безкрайност (Glorfeld, 1995). Паралелният анализ, от своя страна, е извадково-базиран метод, който отчита ефекта на извадковата грешка (Zwick & Velicer, 1986, Hayton, Allen & Scarpello, 2004). Чрез нея се обясняват собствените стойности на латентни променливи, които са равни или по-ниски от „очакваните" собствени стойности на случайно равнище (Horn, 1965; Glorfeld, 1995).

Подобно на всеки перспективен метод, Паралелният анализ търпи развитие в две основни насоки. Една от тях е преходът от симулиране на определен (дори и значителен) брой очаквани собствени стойности при нулева хипотеза за независимост на променливите към генериране на разпределения на очакваните собствени стойности на поредните фактори. По този начин праговите стойности могат вече да не се фикси-

рат върху средната, а върху определени квантил, например медианата (Kabacoff, 2003) или друг квантил, най-често 75-тия, 90-тия, 95-тия или 99-тия от разпределението на поредната собствена стойност, извлечена от случайните данни (Glorfeld, 1995; Kabacoff, 2003; Hayton, Allen & Scarpello, 2004; Ledesma & Valero-Mora, 2007).

Втората насока е свързана с процедурата за симулиране на паралелните съвкупности от данни (корелационни матрици). Вместо генериране на независими променливи от нормално разпределение, А. Буджа и Н. Еюбоглу предлагат процедура за случайна пермутация на емпиричните данни, като броят на възможните различни наредби (пермутации) на множество от n елемента е $n!$ (Buja & Eyuboglu, 1992; Ledesma & Valero-Mora, 2007). По същество това е непараметрична разновидност на паралелния анализ, полезна в случаите, в които разпределението на променливите в популацията се отклонява от нормалното. В същата статия обаче авторите показват, че Паралелният анализ, базиран на генериране на нормални променливи, демонстрира удивителна устойчивост срещу отклоненията от нормалното разпределение, проявявайки определени непараметрични свойства. Поради това изводите, направени въз основа на допускането за нормалност на разпределенията на променливите (т.е. въз основа на „класическия" паралелен анализ), не се различават от тези, направени по метода на пермутациите.

Своите изводи А. Буджа и Н. Еюбоглу правят въз основа на резултатите от симулативно изследване на качествата на Паралелния анализ по метода на случайните пермутации, чрез който се отчита извадковата вариативност. Авторите провеждат експеримент, при който са генерирани „нулеви" собствени стойности (т.е. собствени стойности при нулева хипотеза на независимост на променливите) от 5 разпределения, 4 от които не-Гаусови теоретични разпределения, съществено отклоняващи се от нормалното, и 1 нормално Гаусово разпределение. От всяко от четирите не-Гаусови разпределения са извлечени по 4 извадки с различен обем и брой променливи (т.е. е направен 4 X 4 факторен експеримент), с 1 000 реплики при всяко условие. Резултатите сочат, че „нулевите" квантили (изчислени при горната нулева хипотеза), получени от не-Гаусовите разпределения, се отклоняват от тези, получени от нормалното разпределение, с не повече от ± 0.04 при 99.77% от всички изчислени собствени стойности. Авторите заключават, разликите между Гаусовите и не-Гаусовите разпределения е достатъчно малка и че нулевите разпределения на собствените стойности са във висока степен независими от типа на разпределението, големината на извадката или броя на променливите. Следователно, използването на метода на пермутациите не носи особени предимства и поради това е препоръчително използването на стандартния метод, базиран на нормални променливи, който „се оказва полезен и оправдан" дори и при не-нормални разпределения (Buja & Eyuboglu, 1992, стр. 2).

Паралелният анализ е не само обект на теоретични изследвания, а и на множество конкретни разработки, които да осигурят неговата приложимост. Разработени

са регресионни формули за апроксимиране на очакваните собствени стойности при симулирани корелационни матрици с определен размер (Allen & Hubbard, 1986, Keeling, 2000), статистически таблици с очакваните собствени стойности за данни с различен обем (Buja & Eyuboglu, 1992), макроси за статистическите пакети SAS (O'Connor, 2000; Kabacoff, 2003) и SPSS (O'Connor, 2000), както и редица самостоятелни програми като ViSta-PARAN (Ledesma & Valero-Mora, 2007).

Резултатите от множество съпоставителни изследвания показват, че Паралелният анализ е по-точен и ефективен от конкурентните методи (Zwick & Velicer, 1986; Revelle & Rocklin, 1979; Buja & Eyuboglu, 1992; Glorfeld, 1995; Fabrigar et al., 1999; Kabacoff, 2003; Hayton, Allen & Scarpello, 2004), а за някои автори той е най-доброто решение на проблема за броя на факторите, както показват проведените през последните 15-20 години изследвания по метода Monte Carlo (Reynolds & Kamphaus, 2003; Ledesma & Valero-Mora, 2007).

Така например в свое съпоставително изследване В. Цвик и У. Велисер съпоставят точността на пет от разгледаните по-горе методи за определяне на броя на факторите: Паралелният анализ на Хорн, тестът *MAP* на Велисер, тестът на Кетел, χ^2 тестът на Бартлет и критерият на Кайзер-Гутман (Zwick & Velicer, 1986). Методите са изследвани при различни условия (обем на извадката, брой на променливите и на компонентите) и с различни корелационни матрици. Резултатите убедително сочат, че Паралелният анализ се представя най-добре, тъй като води към коректни решения при 92% от изследваните случаи. При все това авторите отбелязват наличието на слаба тенденция за надценяване на броя на факторите. Забелязани са също прояви на включване в модела и на незначителни фактори. А. Буджа и Н. Еюбоглу също представят емпирични доказателства за тези слабости на *PA*, особено при симулиране на *PFA* (Buja & Eyuboglu, 1992). Решението на този проблем, което А. Буджа и Н. Еюбоглу предлагат, е повишаване на равнището на значимост чрез използване в качеството на прагова стойност на по-висок квантил от разпределението на симулираните собствени стойности – например 5% или 1% в десния му край. Тестът *MAP* се оказва точен при 84% от случаите, с противоположната тенденция за подценяване на техния брой. Добре се представя и тестът на Кетел, който е точен при 57% случаите, с тенденция да надценява броя на факторите, но работи добре, когато латентните променливи са отчетливо структурирани (виж още Fabrigar et al., 1999). χ^2 тестът на Бартлет е коректен само при 30% от случаите, с тенденция за надценяване на броя на факторите, а най-ненадежден е критерият на Кайзер-Гутман с точност едва при 22% от случаите, с доказана "склонност" към съществено надценяване на броя на факторите (виж също Horn, 1965; Buja & Eyuboglu, 1992; Glorfeld, 1995; Fabrigar et al., 1999; Costello & Osborne, 2005).

Подобно изследване с реални данни от личностовия въпросник NEO-PI-R (базиран на теорията за „голямата петорка"), с извадка от 1 000 и. л. провежда и В. Ревел

(Revelle, 2007). Авторът изследва „поведението" на 5 метода за определяне на броя на факторите (Паралелен анализ, тестовете на Кетел и Бартлет, *VSS* и йерархичен клъстърен анализ). Както би могло да се очаква, най-добре възпроизвеждат очакваната 5-факторна структура Паралелният анализ и тестът на Кетел.

Както сочат резултатите от представените изследвания, изводът на Л. Глорфелд, че няма никакви разумни причини да се избере друг метод, освен Паралелния анализ, е напълно основателен (Glorfeld, 1995). Независимо от това изследователите упорито прилагат традиционните методи. Дж. Форд и сътрудници анализират публикациите в три престижни психологически списания за период от 10 години (1975-1984). В общо 152 статии, в които е представено използването на изследователски факторен анализ, най-често използваният метод за определяне на броя на факторите е критерият на Кайзер-Гутман $K1$ (в 21.70% от случаите), при това с нарастваща във времето интензивност. Тестът на Кетел, предшестващ теоретичен модел или възможността за интерпретация са използвани като методи в 11.2% от публикациите, а в 13.8% - някаква комбинация от тези методи. В нито една публикация не е използван методът на паралелния анализ (Ford et al., 1986, по Hayton, Allen & Scarpello, 2004).

Подобен преглед е направен от Л. Фабригар и сътрудници, които анализират две престижни психологически списания за период от 5 години (1991-1995). Справката от наличните 217 статии разкрива картина, подобна на предходната, при това в около 40% от текстовете авторите не са уточнили кой метод за установяване на броя на факторите са използвали (Fabrigar et al., 1999).

2.2.3.4. Постигане на проста факторна структура

Както бе отбелязано по-горе, основната психометрична цел при използването на изследователския факторен анализ е разкриване на структурата на взаимовръзките между променливите и конструиране на скали. Вземайки предвид проблема за определяне на броя на факторите при изграждането на модела, а също и обстоятелството, че различните методи за факторизиране могат да доведат до различаващи се конфигурации, тази обща цел може да бъде преформулирана като „постигане на проста (чиста) факторна структура". Идеята е развита от Л. Л. Търстоун в множество негови публикации (Thurstone, 1934; 1935; 1936). "Търстоуновото разбиране за проста структура – отбелязва А. Йейтс, се опитва да постави един общ набор от ограничения на факторния модел, от който би могло да се очаква да доведе до един научно смислен резултат" (Yates, 1987, стр. 31; Thorndike, 1971; Harman, 1976; Costello & Osborne, 2005).

Концепцията на Л. Л. Търстоун се отнася до третата, заключителна фаза на факторния анализ, при която се извършва ротация на факторите от избрания модел с цел да се достигне до възможно най-простия, но смислен и теоретично съдържателен модел на връзките между манифестираните и латентните променливи. А. Йейтс отбе-

лязва, че „основата на търсенето на „проста структура" е научното вярване в простотата и пестеливостта на естествените процеси (Yates, 1987, стр. 3).

Първоначално Л. Л. Търстоун разработва математически критерий за определяне на адекватна проста структура, който по-нататък развива в три, а по-късно в пет общи условия, предназначени за проверка на единствеността (*uniqueness*) на дадена проста структура, която предстои да бъде приета като факторен модел (Thurstone, 1935, 1947). Тъй като различните автори представят правилата на Л. Л. Търстоун в различаващи се, макар и несъществено, редакции, формулировките по-долу са главно по Х. Харман и А. Йейтс (Harman, 1976, стр. 98; Yates, 1987, стр. 33-37). Използвани са следните обозначения:

p – общ фактор ($1 \leq p \leq m$)
j – зависима променлива ($1 \leq j \leq n$)

1. Всеки ред j от факторната матрица V трябва да съдържа поне едно нулево факторно тегло.

(т.е. всяка наблюдавана променлива трябва да се описва от максимум m - 1 фактора).

2. Всяка колона p от факторната матрица V трябва да съдържа ясно различима група от m линейно независими наблюдавани променливи, чиито факторни тегла v_{jp} са нулеви.

(т.е. всеки фактор трябва да описва максимум n - m наблюдавани променливи).

3. Във всеки две колони от факторната матрица V трябва да има няколко наблюдавани променливи, чиито факторни тегла v_{jp} клонят към нула в едната колона, но не и в другата.

(т.е. които имат високи факторни тегла по единия фактор и близки до нула – по другия).

4. Във всеки две колони от факторната матрица V, голяма част от наблюдаваните променливи трябва да имат клонящи към нула факторни тегла в двете колони. Това правило се отнася за случаите, в които са извлечени 4, 5 или повече фактора.

5. Във всеки две колони на факторната матрица V трябва да има, за предпочитане, малък брой наблюдавани променливи, които имат неклонящи към нула факторни тегла по двете колони.

Х. Харман отбелязва, че условията на Л. Л. Търстоун следва да се прилагат в определени случаи. Според него, има мълчаливо съгласие, че това са случаите, в които факторите са некорелирани (ортогонални), иначе терминът „факторна матрица" би бил напълно неясен (Harman, 1976). А. Йейтс прави уточнението, че само първото от горните условия е свързано пряко с математическия критерий на Л. Л. Търстоун за

проста структура. Останалите четири са начин за оценка на нейната уникалност и стабилност (Yates, 1987).

Концепцията на Л. Л. Търстоун за търсене на проста структура чрез въртене на факторите е широко приета в средите на психометриците, макар и да има своите критици, сред които е и Х. Кайзер, главно поради качествения характер на някои от условията. Р. Торндайк смята, че те водят до некоректно определяне на някои фактори и предлага ревизия на условията. Тя се състои в това, че максимална простота на структурата се постига тогава, когато дадена променлива има определено (високо) тегло само по един от факторите и нулеви тегла по всички останали фактори (Thorndike, 1971).

В по-малко рестриктивна интерпретация, това правило гласи, че проста факторна структура е налице, когато дадена манифестирана променлива има високо тегло по един от факторите и ниски тегла по останалите. Такава простота в структурата може да бъде „постигната" и чрез „отстраняване" от анализа на променливи, които имат факторни тегла под определени граници (например 0.70, 0.50 и т.н). Този подход се прилага въз основа на факта, че факторните тегла представляват регресионни коефициенти на съответните наблюдавани променливи и поради очевидната зависимост между обема на извадката и статистическата значимост на съответното факторно тегло. При фиксирано ниво на значимост (например 0.05), на по-големите по обем извадки биха съответствали по-ниски прагови стойности на факторните тегла.

2.2.3.5. Допускания за прилагането на факторен анализ

(1) Метричност на наблюдаваните променливи

Едно от изискванията на факторния анализ е наблюдаваните променливи да бъдат измерени поне в интервална скала. Изискването следва от използването на матрици на Пиърсъновите корелации (или на ковариациите) като база за факторния анализ, както и от конструирането на факторите като претеглени суми от наблюдаваните променливи (Kim & Mueller, 1978; Kubinger, 2003).

Същевременно много автори отбелязват, че на факторен анализ могат да бъдат подложени и ординални променливи, ако приписаните рангове не изкривяват сериозно метричните особености на латентните променливи, тъй като корелационните коефициенти са устойчиви срещу такива изкривявания (Kim, 1975; Kim & Mueller, 1978). Същото се отнася и за наблюдаваните променливи с органичен брой категории (от ликертов тип), включително и за дихотомични (дихотомизирани) променливи (Knol & Berger, 1991, de Ayala, 2009). Въпросът според авторите не е дали такива променливи могат да бъдат подложени на факторен анализ, а в каква степен по-ниската мощност на неметричните скали би изкривила корелациите между променливите и, следователно, резултатите от анализа. де Аяла обръща специално внимание на това, че дихотомич-

ните данни не винаги създават проблеми и че използването на линеен факторен анализ върху такива и върху рангови данни е напълно приемливо (de Ayala, 2009).

(2) Видове корелации: r на Пиърсън, φ и тетрахоричен коефициент на корелация

За основа на факторния анализ може да бъде взета матрица ковариациите, но обикновено се използва корелационна матрица. В случите, в които наблюдаваните променливи са метрични (измерени в интервална скала), това е матрицата на коефициентите на корелация *r* на Пиърсън.

При дискретни дихотомични данни (номинални променливи, които имат само две възможни стойности) е оправдано използването на подходящи коефициенти на корелация като φ, който представлява специален случай на Пиърсъновия продукт-момент коефициент на корелация (Harman, 1976; Калинов, 2010;). Някои изследователи не само препоръчват използването на φ (Kim & Mueller, 1978), но и демонстрират приложението на тази мярка за взаимовръзка (de Ayala, 2009). Обобщавайки резултатите от свое обширно експериментално изследване на паралелния анализ като метод за определяне на броя на факторите (с използване на φ и на тетрахорични коефициенти), Л. Уенг и Ч. Ченг препоръчват „бъдещите приложения на паралелния анализ на бинарни данни да бъдат извършвани с φ корелации вместо с тетрахорични корелации предвид нестабилното поведение на тетрахоричната корелация" (Weng & Cheng, 2005, стр. 713).

Според Дж. Ким и Ч. Мюлер коефициентът φ е полезен в случаите, в които факторният анализ се използва като средство за общо клъстъризиране на променливите и ако латентните корелации между променливите са умерено силни, по-ниски от 0.60 - 0.70. Причината е, че дихотомизирането на континуалните латентни променливи отслабва корелацията между тях и неговият ефект би бил пренебрежимо малък, ако латентните корелации са по-слаби. При това дихотомизирането не засяга латентната структура на данните, защото факторният анализ се основава на съотношенията между различните корелации. В тези случаи, „ако целта на изследователя е да търси клъстърни модели, използването на факторен анализ може да бъде оправдано" (Kim & Mueller, 1978, стр. 75).

Някои автори предупреждават за възможни усложнения при използването на факторен анализ на дихотомични или дихотомизирани данни, и по-специално при използването на матрици с коефициентите на корелация φ (Torgerson, 1958; Kubinger, 2003; de Ayala, 2009). Един от проблемите е извличането на т. н. „фактори на трудността" (*difficulty factors*), който се състои в това, че в един фактор се обединяват въпроси с еднаква (близка) трудност и следователно при анализа могат да се извлекат толкова фактори, колкото са групите въпроси с еднаква (близка) трудност. Въвличането на трудността като статистика на въпросите в тази проблематика идва от това, че

при една перфектна Гутманова скала трудността на въпроса може да се разглежда като точка на дихотомизиране на съответната латентна променлива. Появата на такива нежелани фактори се дължи на особеност на коефициента φ, съгласно която неговата стойност зависи от маргиналните стойности в таблицата на спрегнатост на съответните два въпроса.

Коефициентът φ може да достигне максималните си стойности (±1.00) само при равенство на маргиналните стойности на съответната променлива (Калинов, 2010). В този случай двата въпроса имат еднаква трудност и биха формирали един фактор, дори и да се характеризират с ниска съгласуваност помежду си. Това са, по думите на К. Кубингер, „изкуствени" фактори, които не отразяват реалната факторна структура на променливите (Kubinger, 2003). Според В. Ревел използването на φ за оценка на латентните корелации би подценило стойностите на r на Пиърсън, предназначени за същата цел (Revelle, 2011). Някои автори обаче показват, че дори и при използване на факторен анализ върху φ, могат да се постигнат резултати, аналогични на резултатите от използването на други методи за определяне на размерността на латентното пространство (de Ayala, 2009).

Начин за избягване на проблемите с коефициента φ при бинарни данни, каквито са резултатите от ТОП на субтестово равнище, е използването на други мерки на взаимовръзките между променливите, различни от Пиърсъновия продукт-момент коефициент на корелация и неговите производни. Такъв подход е използването на матрица на тетрахоричните коефициенти на корелация r_{tet} вместо r или φ, препоръчвано от мнозина изследователи (Torgerson, 1958; Lord, 1980; O'Connor, 2000; Ledesma & Valero-Mora, 2007; de Ayala, 2009; Revelle, 2011).

Наблюдаваните променливи (на ниво тестов въпрос) са формално категориални (дихотомични). Може да се предположи обаче, съгласно латентната теория, че всяка наблюдавана променлива X_j е дихотомична манифестация на съответстваща ѝ латентна променлива X^*_j, т.е. манифестираните променливи не са дихотомични „по природа", а са дихотомизирани. Наблюдаваната чрез коефициента φ корелация между манифестираните променливи е неточна оценка на връзката между техните латентни съответствия. От друга страна, тетрахоричната корелация се разглежда като оценка на корелацията между самите латентни променливи. По-точно, тетрахоричният коефициент r_{tet} между две манифестирани дихотомични променливи е оценка (по метода на максималното правдоподобие) на Пиърсъновия корелационен коефициент, който може да бъде получен, ако съответните две латентни променливи, измерени в метрична скала и разпределени нормално, са наблюдавани пряко (Uebersax, 2000). По този начин факторният анализ на тетрахоричните корелации придобива огромното предимство да борави директно с латентните променливи.

Като недостатък на тетрахоричния коефициент може да се посочи това, че на ниво извадка той се изчислява поотделно за всяка двойка променливи, без съвкуп-

ността от променливи да се третира като многомерно разпределение. Поради това е възможно да се получат и не-положително определени симетрични квадратни матрици.

(3) Нормалност на разпределенията на латентните променливи

Друго фундаментално изискване за прилагането на факторния анализ е променливите в генералната съвкупност (латентните променливи) да имат многомерно нормално разпределение (*multivariate normality*). Това изискване произтича от обстоятелството, че методът борави с корелациите между променливите, които биха били оценени като по-ниски, ако манифестираните променливи произтичат от различни разпределения. Според Р. МакДоналд, само при наличието на това условие „…корелационният коефициент е пряк индекс на степента на взаимовръзка" между двете променливи (McDonald, 1985). Това се отнася не само при използване на Пиърсъновия коефициент r, но и при неговите специални случаи (O'Connor, 2000; Revelle, 2011).

Някои изследователи разпростират изискването за многомерна нормалност върху двата основни метода на факторизиране (PCA и PFA) (McDonald, 1985; O'Connor, 2000; Harris, 2001). Други също определят това изискване като основно, по-специално при методите на максималното правдоподобие (Jöreskog, 1966; Harman, 1976) и на минималните остатъци (Harman, 1976).

Л. Фабригар и сътрудници, които споделят допускането за многомерна нормалност, обвързват метода на факторизиране с вида на разпределението (Fabrigar, Wegener, MacCallum & Strahan 1999). По тяхно мнение, методът на максималното правдоподобие е подходящ за нормални или близки до нормалното разпределения, но ако това допускане е нарушено, те препоръчват използването на някои от другите методи на главните фактори.

Други обаче смятат, че „сам по себе си факторният анализ не изисква такова допускане", т. е., не е необходимо променливите да формират многомерно нормално разпределение (Kim & Mueller, 1978, стр. 77). Авторите посочват обаче, че за прилагане на метода на максималното правдоподобие и свързаните с него тестове за значимост това допускане е валидно. Х. Харман също приема, че при използване на продукт-момент корелации не е необходимо да се прави допускане за нормалност. Но за да се направи по-смислена интерпретация на получените резултати, авторът разглежда като „желано" допускането за двумерно нормално разпределение на променливите (Harman, 1976, стр. 24).

Що се отнася до тетрахоричния коефициент на корелация, изискването за двумерно нормално разпределение на съответните (двойки) латентни променливи е безспорно (Fabrigar, Wegener, MacCallum & Strahan 1999; Harman, 1976; Kim & Mueller, 1978). Х. Харман обаче предупреждава, че все още не е разработен добър математи-

чески модел за изчисляване на тетрахоричния корелационен коефициент от данни, които се характеризират с многомерно нормално разпределение. Поради това има възможност матрицата от такива коефициенти да не бъде консистентна и съответно неподходяща за факторен анализ (Harman, 1976). Като цяло, обаче, „последствията от нарушаването на това допускане не са много ясни" (Kim & Mueller, 1978, стр. 77).

2.3. Резултати

Случай 1. Факторна структура на ТОП на равнище субтест (компоненти на анализа са въпросите във всеки субтест, к=10)

2.3.1.1. Генериране на корелационните матрици

Първата стъпка при факторния анализ на данните е формиране на матриците на тетрахоричните корелации между въпросите в отделните субтестове. Изчислителните трудности при определянето на тези коефициенти се разглеждат като една от основните пречки пред използването на тази мярка на взаимовръзка между променливите. Наличието на различни статистически софтуерни пакети за компютърна обработка на данни обаче премахва тази пречка. За целта бе използван модулът *Reliability & Item analysis* от статистическия пакет STATISTICA, чрез който могат да бъдат изчислени както стандартните Пиърсънови корелационни таблици, така и тетрахоричните. Както бе отбелязано, линейният Пиърсънов коефициент е подходящ за интервални данни, но не и за дихотомизирани данни. Като пример в приложение 5 са представени две корелационни таблици – със „стандартните" линейни коефициенти на Пиърсън и с тетрахорични коефициенти, изчислени за данните от вариант 134, субтест 3. *История*. Може лесно да се забележи разликата в равнищата на двата вида коефициенти - като цяло тетрахоричните коефициенти между отделните двойки въпроси са значително по-високи от съответните им Пиърсънови коефициенти.

По същество на анализът на факторната структура на всички избрани варианти на ТОП на субтестово равнище в настоящото изследване предполага формирането на серия от 120 корелационни матрици (12 теста Х 10 субтеста). Тук ще бъдат представени резултатите от анализа на три от вариантите – 134, 141 и 171, които предизвикват интерес поради това, че при предварителното изследване на корелационните матрици два от тях се очертаха като „гранични" случаи. Вариант 134 се характеризира с най-високи корелационни равнища, вариант 141 – с най-ниски, а вариант 171 заема средищна позиция между тях.

2.3.1.2. Извличане на първоначалните (незавъртени) факторни конфигурации

Като обща стратегия, последователното извличане на факторите се основава на ротация на изходното пространство на променливите, чиято цел е максимизиране на

дисперсията в „новите" променливи (фактори) и същевременно минимизиране на дисперсията в областите между тях (*variance maximizing*). След като първият фактор е конституиран така, че да обясни възможно най-голяма част от споделената дисперсия, алгоритъмът продължава с извличането на следващия фактор, който максимизира останалата дисперсия и така нататък, докато бъде обхванатата и обяснена общата дисперсия. Тъй като всеки следващ фактор обхваща дисперсията, която не е обхванатата от предходния/ предходните, последователните фактори са (а) независими един от друг (некорелирани, ортогонални) и (б) с намаляваща обяснителна сила.

Факторните анализи по метода на главните оси на всички корелационни матрици са направени при такава първоначална конфигурация на факторните модели, която предполага наличието на 10-факторна структура, т. е. при допускане, че броят на факторите е равен на броя на зависимите променливи (тестови въпроси), при фиксирана минимална собствена стойност на факторите $\lambda_{Fi} = 0.00$. Основанията за приемане на такава стартова конфигурация са, че теоретично е възможно компетентността на изпитваните лица по даден въпрос да е напълно независима от компетентността им по останалите въпроси в съответния субтест. С други думи, допускаме, че е възможно отговорите на и. л. на всеки въпрос да кореспондират с отделен латентен фактор и че тези фактори са независими помежду си. От техническа гледна точка тази стартова конфигурация отразява преднамерено търсеното съчетание от максимален брой фактори при минимално равнище на собствените им стойности, което дава възможност да бъдат експлицирани всички „налични" фактори, дори и най-слабите. В допълнение, необходимостта от прилагане на такава първоначална конфигурация се налага и от използването на паралелния анализ като метод за определяне на броя на факторите (O'Connor, 2000). Пълните резултати от направените факторни анализи при избраната стартова конфигурация, както и резултатите от последващите допълнителни анализи, са представени в приложение 6.

Необходимо е да се отбележи, че факторният анализ по метода на главните оси, при избраната първоначална конфигурация, бе приложен успешно върху всички матрици на тетрахоричните корелации, с изключение на тази от субтест 4. *География,* вариант 134. Причината е, че при факторния анализ се използва обратната корелационна матрица, чиито елементи в главния диагонал, определяни като „фактори за увеличаване на дисперсията" (*variance inflation factors*), се изчисляват по формулата:

$$VIF_j = \left(1 - R_j^2\right)^{-1} \qquad (61)$$

където:

VIF_j – фактор за увеличаване на дисперсията на j-тата променлива

R_j^2 - коефициент на множествена корелация (детерминация) на j -тата променлива с другите променливи

Ако дадена променлива (тестов въпрос) x_j не корелира с останалите променливи (R_j = 0.00), тогава стойността на съответния фактор за увеличаване на дисперсията VIF_j е равен на 1.00. От горното уравнение следва още, че при наличие на коефициенти на множествена корелация, равни на 1.00, съответният фактор за увеличаване на дисперсията не може да бъде изчислен. Този негативен ефект се получава в случаите, когато между (някои от) променливите се наблюдава (мулти)колинеарност, т. е. висока корелация, която може да се тълкува като сигнал, че част от тези променливи са „излишни". Поради това факторите за увеличаване на дисперсията се разглеждат и като мярка за оценка на мултиколинеарността между променливите.

Изследователите предлагат различни подходи за решаване на този проблем. М. Пет и сътрудници предлагат 4 такива решения, отнасящи се до проверка на данните за дублиране/сходство по редове (т. е. между и. л.) или по колони (т. е. между променливите). Особено важна е проверката за наличие на високи корелации между айтемите. Според авторите, ако в съответната матрица има корелационни коефициенти, по-високи от 0.80, от анализа следва да бъдат извадени един или повече от тези айтеми (Pett, Lackey & Sullivan, 2003, стр. 72). При вариант 134, субтест 4. *География* между два от въпросите с поредни номера 33 и 34 бе наблюдавана корелация r_{tet}=1.00. След изваждане на въпрос 33, бе формирана на нова корелационна матрица от останалите 9 въпроса, която бе подложена успешно на факторен анализ.

Най-напред ще направим общо представяне на първоначалните резултати от приложения изследователски факторен анализ по метода на главните оси за всички изследвани субтестове по данните от приложение 6. Броят на първоначално извлечените фактори при различните субтестове варира между 4 и 7, като по-често срещаните конфигурации включват 5 фактора (при 14 субтеста, 46.67% от всички) или 6 фактора (при 12 субтеста, 40.00% от всички). Информация за останалите конфигурации може да бъде извлечена от следващата таблица.

Таблица 2. Брой на първоначално извлечените фактори

Брой на първоначално извлечените фактори	Честота (брой субтестове)	Процент от всички субтестове
4	2	6.67
5	14	46.67
6	12	40.00
7	2	6.67

Най-важната мярка за оценка на вариацията в множество от наблюдавани променливи, която се дължи на (и може да бъде обяснена чрез) даден фактор, е неговата собствена стойност (*eigenvalue*). Тази статистика е основен индикатор за значимостта, за обяснителната сила на съответния фактор. Поради спецификата на алгоритъма за

последователно извличане на факторите, получените първоначални факторни конфигурации се характеризират с голям брой, но неравностойни фактори, всеки следващ от които обяснява все по-малка част от наблюдаваната дисперсия. Нещо повече, поради специфичния подход на *PFA* за анализиране само на общата дисперсия на променливите, която се дължи на общите фактори и която те споделят помежду си, извлечените фактори обикновено не обясняват цялата дисперсия.

Като примери по-надолу ще разгледаме резултатите от анализа на два субтеста, съответно субтест 7. *Химия* от вариант 141 и субтест 5. *Математика* от вариант 171, представени на следващата таблица.

Таблица 3. Резултати от факторния анализ (метод на главните оси) на данните от субтест 7. *Химия,* вариант 141 и субтест 5. *Математика*, вариант 171

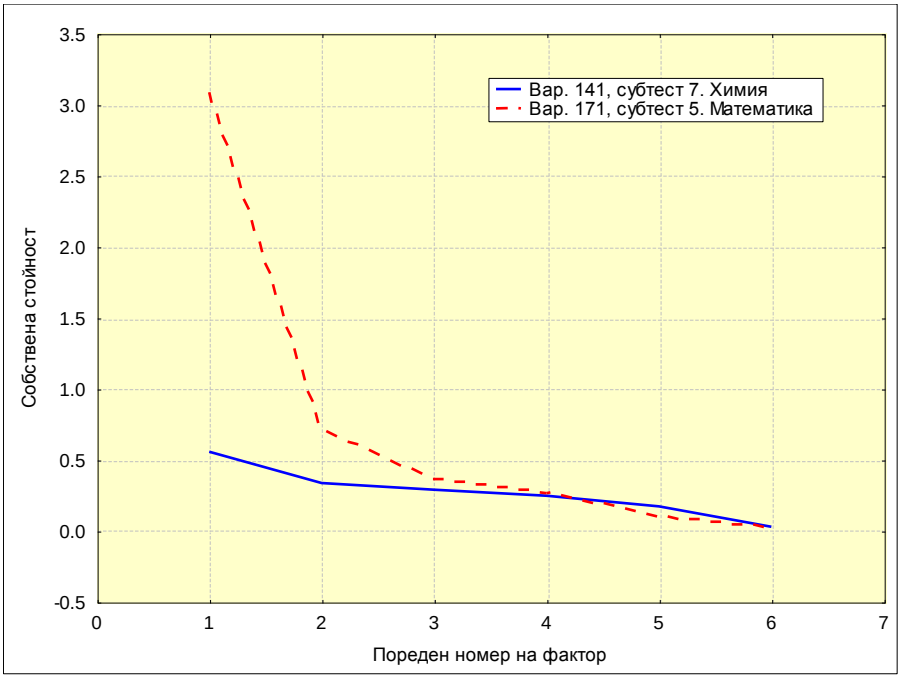
Източник на данни	**Статистики на реалните данни**				**Статистики на симулираните данни**	
	пореден номер на фактор	**собствена стойност**	**% от цялата дисперсия**	**кумул. % от цялата дисперсия**	**средни ст-ти**	**95-ти процентил**
1	2	3	4	5	6	7
вариант 141 субтест 7. Химия	F1	0.556	5.562	5.562	1.186	1.236
	F2	0.336	3.360	8.922	1.130	1.166
	F3	0.289	2.888	11.811	1.088	1.119
	F4	0.247	2.465	14.276	1.049	1.075
	F5	0.172	1.718	15.994	1.014	1.039
	F6	0.028	0.282	16.276	0.979	1.002
	F7	-	-	-	0.946	0.970
	F8	-	-	-	0.911	0.937
	F9	-	-	-	0.872	0.903
	F10	-	-	-	0.826	0.863
вариант 171 субтест 5. Математика	F1	3.086	30.862	30.862	1.174	1.220
	F2	0.715	7.146	38.008	1.122	1.157
	F3	0.373	3.732	41.740	1.082	1.111
	F4	0.265	2.646	44.386	1.045	1.070
	F5	0.103	1.027	45.413	1.014	1.037
	F6	0.032	0.321	45.734	0.982	1.005
	F7	-	-	-	0.949	0.973
	F8	-	-	-	0.916	0.941
	F9	-	-	-	0.880	0.909
	F10	-	-	-	0.836	0.872

От една страна, по отношение на своята факторна структура, тези субтестове са типични представители на изследваните субтестови корелационни матрици. От друга страна, те представляват не само две различни образователни области, но и различ-

ни, в известен смисъл контрастни резултати, тъй като двата субтеста са с най-ниска и съответно най-висока собствена стойност на първия фактор сред всички анализирани субтестове.

В горната таблица, освен собствените стойности на първоначално извлечените фактори (колона 3), са представени и техните дялове (в проценти) от цялата (*total*) дисперсия на променливите (колона 4), както и кумулативният процент на дисперсията, обяснена от съответния брой фактори. Съгласно получените резултати, общата дисперсия при разглежданите два субтеста може да бъде обяснена с факторни структури, включващи 6 фактора, като най-силен е фактор 1 от раздел 5. *Математика* със собствена стойност 3.086, който обяснява 30.862% от цялата дисперсия в този раздел, а най-слаб – фактор 6 от раздел 7. *Химия* със собствена стойност 0.028, на който се дължи едва 0.282% от цялата дисперсия в раздела. Останалите фактори се разполагат между тези гранични стойности. Ако съпоставим собствените стойности на първите по ред фактори в двата субтеста, бихме могли да отбележим очевидно ниската собствена стойност на първия фактор от субтеста по химия (0.556), който обяснява също така нисък дял от общата дисперсия (5.562%) и относително високата собствена стойност на първия фактор от субтеста по математика, на който се дължи и относително голям дял от общата дисперсия на въпросите. Друга особеност е сравнително малката разлика между собствените стойности на факторите от субтеста по химия и относително голямата разлика между собствените стойности на (някои от) факторите от субтеста по математика. Тази особеност е илюстрирана на следващата графика.

Фигура 12. Собствени стойности на факторите от, субтест 7. *Химия* от вариант 141 и субтест 5. *Математика* от вариант 171



Профилът на субтеста по математика е силно скосен, особено в частта наляво от втория фактор. В този профил се откроява ярко голямата собствената стойност на първия фактор, чието равнище е значително по-високо от тези на останалите фактори, включително и от втория по ред. За разлика от него, профилът на субтеста по химия е почти хоризонтален, с ниска собствена стойност на първия фактор и също така ниски, намаляващи с малка стъпка собствени стойности на останалите фактори.

По-силният контраст в равнищата на собствените стойности в субтеста по математика, особено между първия и втория фактор, предоставя много по-добра основа за интерпретация на неговата факторна конфигурация, отколкото при субтеста по химия. Както бе отбелязано, субтестовете по математика и химия, представени на горната графика, са съответно с най-висока и най-ниска собствена стойност на първия фактор, поради което първоначалните факторни конфигурации на останалите субтестове са разположени някъде между тези две.

Независимо от големия брой на факторите в незавъртените начални решения (от 4 до 7 при 10 манифестирани променливи), както и от спецификата на анализа на главните фактори, предназначен за обяснение на споделената, а не на цялата дисперсия, тяхната съвкупна обяснителна сила не е висока. Кумулативният процент от цялата дисперсия, обяснена от всички фактори (кол. 5) при вариант 141, субтест 7. *Химия* е 16.28%, а при вариант 171, субтест 5. *Математика* – 45.73%. Останалата, по-голяма част от наблюдаваната дисперсия, съгласно уравнение (58), се дължи на специфични характеристики на въпросите и/или на грешката на измерване. Впрочем, делът на дисперсията в съвкупността от анализирани субтестове, обяснена чрез първоначално извлечените факторни конфигурации, варира в границите, установени при тези два субтеста (16.28% - 45.73%).

На следващата таблица е представено групирано разпределение на дяловете от цялата дисперсия при всички анализирани субтестове, обяснена чрез съответната първоначална факторна конфигурация.

Таблица 4. Разпределение на дяловете от цялата дисперсия, обяснени от първоначалните (незавъртени) факторни конфигурации

Дял от цялата дисперсия (в проценти)	Брой субтестове	Процент от всички субтестове	Кумулативен процент
15.00<x<=20.00	5	16.67	16.67
20.00<x<=25.00	13	43.33	60.00
25.00<x<=30.00	3	10.00	70.00
30.00<x<=35.00	3	10.00	80.00
35.00<x<=40.00	3	10.00	90.00
40.00<x<=45.00	2	6.67	96.67
45.00<x<=50.00	1	3.33	100.00

От данните в горната таблица става ясно, че първоначалните конфигурации в преобладаващата част от субтестовете обясняват относително малка част от цялата наблюдавана дисперсия. В 60% от субтестовете тази част е до 25%, като най-често срещаният дял е между 20.00% и 25.00%, който се наблюдава при 13 субтеста (43.33% от всички).

2.3.1.3. Определяне на факторните модели

Като основен метод за решаване на проблема за определяне на броя на факторите, които следва да бъдат включени в модела, в настоящото изследване бе предпочетен паралелният анализ на Дж. Хорн, представен подробно по-горе в текста. Паралелният анализ бе извършен с помощта на макрос, разработен за статистическия пакет SPSS от Б. О'Конър (O'Connor, 2000). Изчисленията на „очакваните" собствени стойности при всички субтестове са направени при следните стойности на основните параметри:

(1) брой на случаите – при вариант 134 n = 713, при вариант 141 n = 730, при вариант 171 n = 830

(2) брой на променливите – 10

(3) брой на симулираните извадки – 1 000

(4) процентил – 95

(5) тип на анализа -1

Очакваните собствените стойности са изчислени при нулева хипотеза за независимост (липса на корелация) между променливите, генерирани от симулирана, нормално разпределена генерална съвкупност. Основните параметри при конфигуриране на паралелния анализ на субтестово равнище имат следното значение:

(1) Броят на случаите във всяка симулирана променлива е равен на обема на извадката при съответния реален субтест.

(2) Броят на генерираните променливи за всеки симулиран субтест е равен на броя на въпросите в съответния реален субтест (за всички субтестове от ТОП този брой е еднакъв).

(3) Брой на генерираните извадки, за всяка от които се изчисляват „очакваните" собствени стойности на 10 поредни фактора. Дж. Хорн препоръчва броят на извадките с рандомизирани данни да бъде достатъчно голям (Horn 1965). Въпреки че не е установен някакъв стандарт, общото правило е, че колкото по-голям е техният брой, толкова по-точни ще бъдат резултатите. Правени са изследвания със 100, 500 и 1 000 случайни извадки, без да се наблюдават значими различия в резултатите. В настоящото изследване също беше направено тестване с различен брой извадки. Не се наблюдават съществени разлики между изчислените собствени стойности, но се забелязва слаба тенденция равнищата им да намаляват. Зададеният брой от 1 000 извадки не предизвиква изчислителни проблеми и гарантира достатъчни по обем вариационни

редове от симулирани собствени стойности, от които да се определят референтните стойности.

(4) 95-тият процентил на разпределението на собствените стойности за всеки пореден симулиран фактор. В настоящото изследване 95-тият процентил е приет като критерийна (референтна) стойност за вземане на решение дали съответният фактор да бъде включен в модела. Освен тези стойности, са изчислени и средните стойности на симулираните собствени стойности. Двете поредици от средни собствени стойности и 95-процентилни стойности, извлечени от разпределенията на 1 000 симулирани извадки, са представени в кол. 6 и 7 на таблица 3 и съответно в таблицата, дадена в приложение 6.

Както беше отбелязано по-горе, в първоначалния модел на метода Дж. Хорн препоръчва за референтни да се използват средните симулирани собствени стойности. Това е подобно на установяване на критична стойност от 0.50 на грешките от I род (Glorfeld, 1995). Авторът препоръчва като по-консервативна мярка използването на 95-тия процентил на разпределенията на симулираните собствени стойности, който съответства на критична стойност от 0.05 на грешките от I род. Някои автори препоръчват паралелният анализ да се използва успоредно с графичния тест на Кетел, който дава възможност за визуално съпоставяне на действителните със симулираните собствени стойности, дори и с критерия на Кайзер $K1$ (Fabrigar et al., 1999; Ford et al., 1986; Hayton et al., 2004). Същите автори отбелязват, че поради очевидната взаимозависимост между собствените стойности на факторите, наличието на силен първи фактор може да доведе до „отслабване" на останалите фактори, т. е. до подценяване на техния брой.

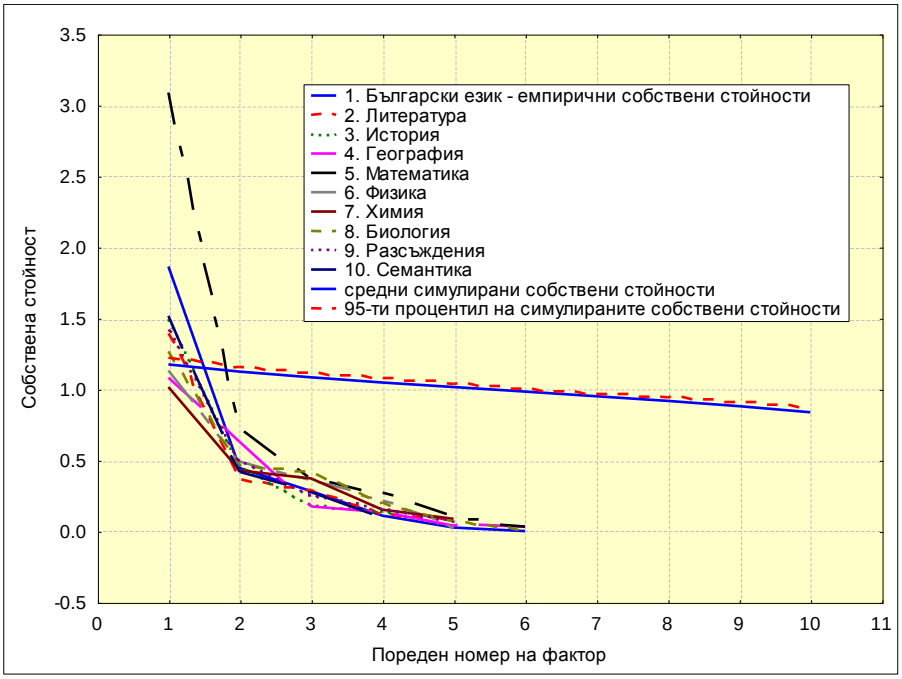
(5) Типът на факторния анализ на симулираните данни. В случая е определен анализ на главни компоненти, който е обичаен метод за определяне на броя на факторите при факторния анализ на главни оси. Впрочем, мненията на изследователите по този въпрос се различават (O'Connor, 2000). Привърженик на подхода за определяне на оптималния брой на общите фактори чрез собствени стойности, получени по метода на главните компоненти, е Р. Кетел, авторът на графичния тест (Cattell & Vogelmann, 1977; Cattell, 1978; O'Connor, 2000).

Определянето на факторния модел (по-скоро – решението на проблема за броя на факторите) се извършва чрез съпоставяне на емпиричната собствена стойност на всеки фактор, определена върху реалните данни, с критерийната му стойност, определена върху симулираните данни.

Прилагането на паралелния анализ като метод за определяне на броя на факторите в съответните субтестови конфигурации доведе до съществени изменения спрямо първоначалните им варианти, описани в предходната част на текста. По-голяма част от латентните модели, които следва да се определят въз основа на избраната процедура, са едномерни, включващи един фактор.

Като илюстрация на механизмите на вземане на решения за определяне на броя на факторите, основани на паралелния анализ, на следващата графика е представена диаграмата на емпиричните и симулираните собствени стойности на всички субтестове от вариант 134. По същество този тип изображения представляват развитие на графичният тест на Р. Кетел, към който са добавени и резултатите от съответните паралелни тестове - очакваните средни и 95-процентилни симулирани собствени стойности. Забелязва се, че средните на симулираните собствени стойности са малко по-ниски от съответните 95-процентилни стойности, което показва, че разпределенията на симулираните собствени стойности за всеки пореден фактор имат много малка дисперсия.

Фигура 13. Приложение на паралелния анализ върху данните от факторния анализ на субтестовете от вариант 171



Профилът на наблюдаваните собствени стойности (виж легендата) е типичен компонент от графичния тест на Р. Кетел. Всяка негова точка представя големината на собствената стойност на съответния пореден фактор. На тази графика се забелязват открояващите се първи по ред фактори на отделните субтестови конфигурации, всички със собствени стойности, по-високи от 1.00, и поредица от значително по-слаби следващи по ред фактори с ниски, намаляващи собствени стойности („факторни отломки"), вариращи от 0.71 (при субтест 5. *Математика*) до 0.00 (при субтест 1. *Български език*). Дори и вторите по ред фактори се характеризират със собствени стойности, които са на равнища около 0.50, докато тези на последните по ред фактори клонят към 0.00,

При паралелния анализ обаче оценката на важността на отделните фактори се

извършва чрез съпоставяне на наблюдаваната собствена стойност на всеки фактор с критерийната (референтна) стойност, за каквато в настоящото изследване е избран 95-тия процентил от разпределението на случайните собствени стойности при дадената нулева хипотеза.

Данните от таблицата в приложение 6, илюстрирани на горната графика, показват, че само 7 от субтестовете на вариант 171 имат първи фактор с наблюдавана собствена стойност, по-голяма от съответната референтна стойности, т. е. която надхвърля тази, извлечена на случайно равнище. Първите по ред фактори от конфигурациите на останалите три субтеста имат собствени стойности, които могат да се разглеждат като несъществени, получени на случайно равнище. Подобна е картината на резултатите от факторните анализи на другите два варианта на теста. На следващата таблица е представено обобщение на получените резултати. Със знак (+) са маркирани субтестовете, чийто първи фактор е значим, а с (-) са отбелязани тези, при които собствената стойност на първия фактор е на случайно равнище.

Таблица 5. Субтестове със значим първи фактор

Субтест	Вариант 134	Вариант 141	Вариант 171
1. Български език	+	+	+
2. Литература	+	-	+
3. История	+	-	+
4. География	+	-	-
5. Математика	+	+	+
6. Физика	+	-	-
7. Химия	-	-	-
8. Биология	-	-	+
9. Разсъждения	+	+	+
10. Семантика	+	+	+

Данните от таблицата разкриват една интересна тенденция. Някои от субтестовете се отличават със значими собствени стойности на първия фактор при всички анализирани тестови варианти. Това са 1. *Български език*, 5. *Математика*, 9. *Разсъждения* и 10. *Семантика*. Собствените стойности на първите фактори на останалите шест субтеста се колебаят около праговата симулирана собствена стойност, позиционирайки се малко над или под нея. Особено интересен е субтест 7. *Химия*, чийто първи фактор не успява да надхвърли случайното равнище при нито един от тестовите варианти. Като цяло, 11 от общо 30 анализирани субтеста (над 1/3 от всички) се характеризират с незначими първи фактори. Тези особености повдигат въпроса за устойчивостта на факторните модели, който ще бъде дискутиран по-нататък. По-важно е да се отбележи, че въпреки наличието на един относително силен първи фактор във всяка латентна структура, допускането за едномерност на тези структури на субтестово равнище

може да бъде подложено на съмнение, поне за онези субтестове, които се характеризират с първи фактори с незначими собствени стойности.

Съществото изменение в броя на факторите след прилагането на процедурата на паралелния анализ доведе и до съществена промяна в дяловете от цялата дисперсия, които могат да бъдат обяснени с приетите еднофакторни модели. Беше отбелязано, че факторите, следващи първия по ред, при всички субтестовите конфигурации се отличават с ниски собствени стойности и съответно имат скромен принос към цялата дисперсия на въпросите.

Таблица 6. Разпределение на дяловете от цялата дисперсия, обяснени от първоначалните и финалните факторни конфигурации

Дял от цялата дисперсия (в проценти)	Първоначални конфигурации		Финални конфигурации	
	брой субтестове	процент от всички субтестове	брой субтестове	процент от всички субтестове
5.00<x<=10.00	-	-	6	20.00
10.00<x<=15.00	-	-	12	40.00
15.00<x<=20.00	5	16.67	3	10.00
20.00<x<=25.00	13	43.33	7	23.33
25.00<x<=30.00	3	10.00	1	3.33
30.00<x<=35.00	3	10.00	1	3.33
35.00<x<=40.00	3	10.00	-	-
40.00<x<=45.00	2	6.67	-	-
45.00<x<=50.00	1	3.33	-	-

Въпреки това техният кумулативен дял се доближава до собствения принос на първия фактор, я в някои случаи го надхвърля. В горната таблица е направена съпоставка на дяловете от цялата дисперсия, която може да бъде обяснена с първоначалните и финалните конфигурации.

В таблицата се наблюдава повсеместно движение надолу, в посока към намаляване на дяловете от цялата дисперсия, която може да бъде обяснена с еднофакторните модели. Докато с първоначалните многофакторни конфигурации може да бъде обяснена до 50% от нея (по-точно - 45.73% при субтест 5. *Математика* от вариант 171), то с еднофакторните модели този максимум е сведен до 35% (30.86% при същия субтест). Типичният дял от общата дисперсия, генерирана от общия фактор, е 10.00 - 15.00% и това е характерно за 12 субтеста (40% от всички). Смущаващо е, че това е горната граница на обяснителната сила на еднофакторните конфигурации в над половината от субтестовите. Сред тях са, разбира се, и субтестовете с номера 2. *Литература* (8.61% от цялата дисперсия), 4. *География* (9.41%), 6. *Физика* (8.42%), 7. *Химия* (5.565) и 8. *Биология* (8.27%), всички от вариант 141. Това са същите онези варианти

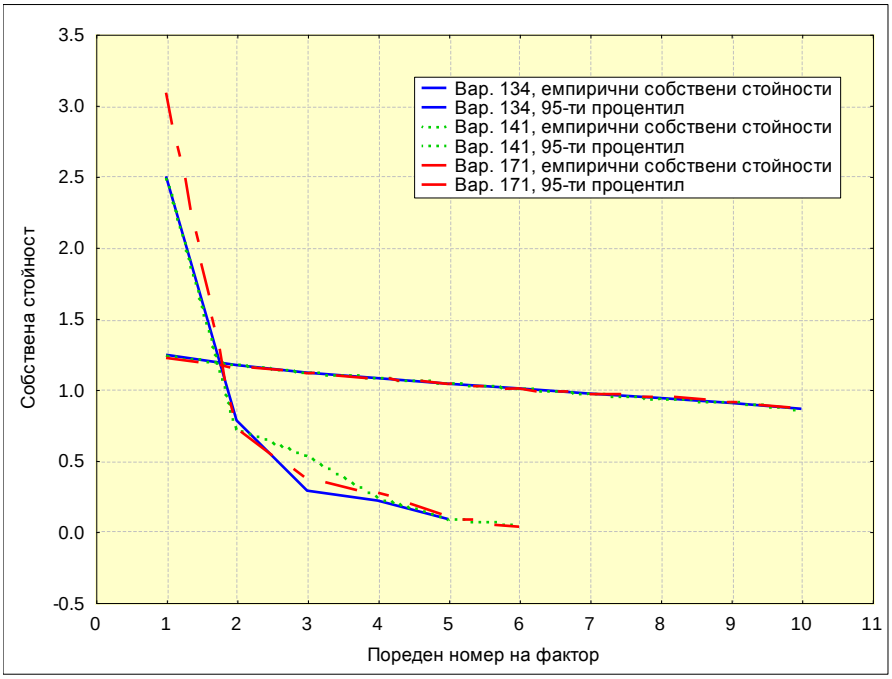
от таблица 5, чиито факторни конфигурации се отличават с незначителен първи фактор.

2.3.1.4. Устойчивост на първоначалните факторни конфигурации

Данните от предходните анализа показват, че профилите на собствените стойности на различните субтестове се различават, както се различават и техните факторни конфигурации по отношение на броя на факторите. Ето защо, независимо от това, че за по-голяма част от субтестовете може да се приеме еднофакторна латентна структура, би било интересно да се проследи дали се наблюдава определена устойчивост (повторяемост) на факторните конфигурации на едни и същи субтестове в различните тестови варианти. Разглеждайки ги, съгласно базовия модел на СТТ, за τ-конгенерични, те се различават по отношение на своето съдържание (включват различни извадки от въпроси), така и по отношение на индивидите (включват различни извадки от и. л.). И при тази плоскост на анализа устойчивостта, повторяемостта на факторните конфигурации е по-скоро изключение, отколкото правило. Нека да разгледаме два типични случая, които илюстрират тази особеност.

На следващата графика са представени профилите на собствените стойности на факторите на раздел 5. *Математика* от тестовите варианти 134, 141 и 171. Тя илюстрира относителната устойчивост на факторните конфигурации, които, без съмнение, при този субтест са твърде сходни.

Фигура 14. Собствени стойности на факторите от раздел 5. *Математика*, варианти 134, 141 и 171



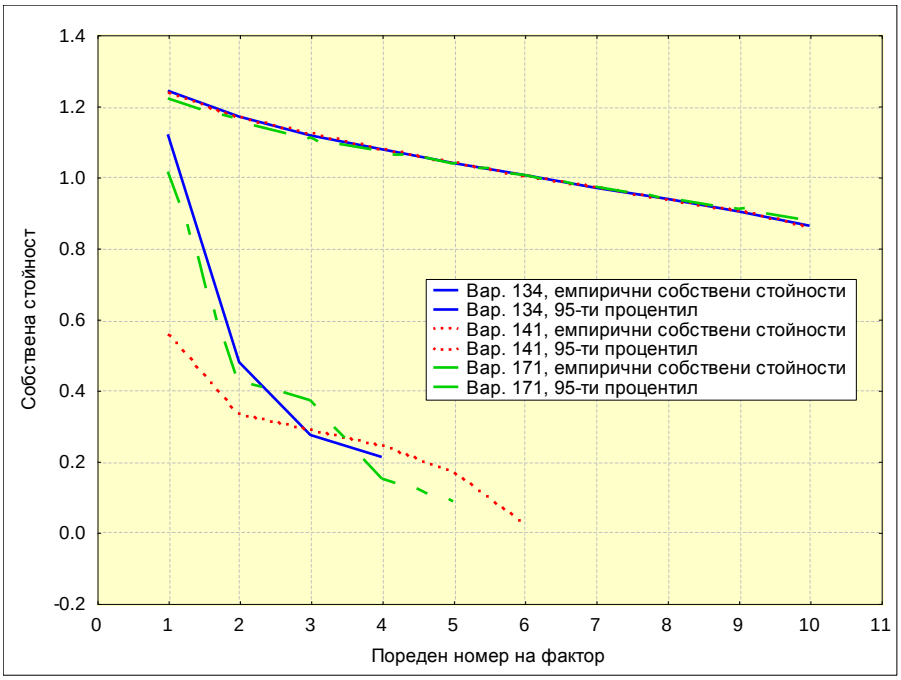
Факторните профили на субтеста по математика се характеризират с 5 - 6 фак-

тора. Ярко изразени, доминиращи са първите по ред фактори, които при трите субтеста имат собствени стойности, далеч по-високи от праговия 95-ти процентил.

Следващите по ред фактори са със значително по-ниски, слабо различаващи се една от друга собствени стойности. Собствените стойности на първите фактори при два от субтестовете (от варианти 134 и 141) са много близки, почти съвпадащи (съответно 2.50 и 2.49); малко по-висока, но близка до първите две, е собствената стойност на първия фактор на субтеста от вариант 171 (3.09). Близки, почти съвпадащи са собствените стойности при следващите по ред фактори, с изключение на третия.

По-често обаче се наблюдава съществена диференциация между профилите на латентните структури на един и същи субтест, както е видно от следващата графика. На нея са представени собствените стойности на факторите от раздел 7. *Химия* от същите три варианта на теста. Латентните конфигурации при този субтест се характеризират с по-голямо разнообразие както по отношение на броя на факторите (4 – 6), така и по отношение на големините на собствените им стойности. Докато между субтестовете от варианти 134 и 171 може да се говори за определено сходство, то профилът на субтеста от вариант 141 е напълно различен от тях. Субтестовите профили от първите два варианта се характеризират с по-ярко изразен първи фактор, и със скокообразно намаляващи собствени стойности на следващите по ред фактори. При субтеста от вариант 141 се наблюдава почти хоризонтална факторна конфигурация с голям брой слабо различаващи се фактори.

Фигура 15. Собствени стойности на факторите от раздел 7. *Химия*, варианти 134, 141 и 171



Собствените стойности на факторите с един и същи пореден номер от различ-

ните профили също се различават значително. Независимо от откроявящите се първи фактори при субтестовете на два от вариантите, нито един от тях не надхвърля граничния 95-ти процентил.

Наличието на множество фактори със слаби, приблизително еднакви собствени стойности не следва да се интерпретира непременно като недостатък на съответната латентна структура. Като недостатък обаче трябва да се разглежда липсата на значими, доминиращи първи фактори, съответно ниските равнища на техните собствени стойности, които за субтеста по химия варират в границите 0.56 (вариант 141) до 1.12 (вариант 134).

2.3.1.5. Анализ на факторните тегла на въпросите от финалните факторни решения. Постигане на прости факторни структури

Ключови за интерпретацията на извлечените фактори са факторните тегла на променливите. Чрез тях се прави оценка на силата на връзката между факторите и съответните променливи. Факторните тегла на променливите могат да се интерпретират като коефициенти на корелация между тези променливи и съответния латентен фактор (Калинов, 2010) или като техни регресионни коефициенти върху фактора/ факторите. От друга страна, факторните тегла при анализа на главни компоненти се интерпретират като проява на силата на въздействието, което съответният фактор оказва върху променливите. Факторните тегла се използват за определяне на степента на „принадлежност" на променливите към съответния фактор и, не на последно място, за неговата интерпретация. В общия случай, дадена променлива може да корелира повече или по-малко с всеки фактор. Нека да разгледаме факторните тегла на тестовите въпроси в субтестовете, които конституират вариант 134.

Спираме се на този вариант, защото, като цяло, той се характеризира с най-високи собствени стойности на първия фактор, т.е. при него едномерността на латентното пространство е най-ясно изразена. Факторните матрици на субтестовете от този вариант са представени в следващата таблица, а на варианти 141 и 171 – в приложение 7. Нека да се фокусираме върху анализа на особеностите на факторните тегла на въпросите.

Въпросите в отделните субтестове имат определени тегла, различни от нула, по съответния субтестов фактор, т. е. всяка променлива корелира, макар и в различна степен, с фактора от съответната латентна структура. Може да се отбележи, че като цяло факторните тегла на въпросите (по абсолютна стойност) не са особено високи. Нещо повече, забелязват се немалко въпроси, които имат изключително ниски тегла по съответния фактор. Такива са например въпрос 8 от субтеста по история (с факторно тегло 0.071), въпрос 1 от субтеста по физика (0.006), въпрос 4 от субтеста по химия (-0.046) и др. Разбира се, такива факторни тегла със стойности, близки до нула, се наблюдават и във факторните матрици на останалите тестови варианти.

Таблица 7. Факторни матрици на въпросите в субтестовете от вариант 134

Номер на въпрос	1. Български език	2. Литература	3. История	4. География	5. Математика	6. Физика	7. Химия	8. Биология	9. разсъждения	10. Семантика
1.	0.514	0.289	0.428	0.396	0.455	0.006	0.509	0.386	0.669	0.542
2.	0.465	0.416	-0.439	0.421	0.681	0.332	0.373	0.136	0.611	0.423
3.	0.636	0.468	0.186	-	0.518	0.566	0.520	0.095	0.473	0.294
4.	0.570	-0.129	0.222	0.272	0.542	0.054	-0.046	0.116	0.615	0.431
5.	0.323	0.487	0.565	0.408	0.728	0.380	0.074	-0.073	0.499	0.442
6.	0.658	0.194	0.536	0.136	0.467	0.122	0.473	0.414	0.464	0.455
7.	0.392	0.496	0.426	0.250	0.015	0.358	-0.180	0.376	0.516	0.722
8.	0.591	0.435	0.071	0.396	0.450	0.517	0.395	0.393	0.431	0.631
9.	0.332	0.287	0.685	0.438	0.494	0.303	0.150	0.361	0.541	0.286
10.	0.322	0.322	0.426	0.513	0.264	0.434	0.091	0.280	0.507	0.422
обяснена дисперсия	2.464	1.389	1.903	1.267	2.498	1.267	1.119	0.872	2.889	2.327
% от цялата дисперсия	24.60	13.90	19.00	14.10	25.00	12.70	11.20	8.70	28.9	23.30

*Забележки: *Обяснена дисперсия – собствена стойност на съответния фактор*

*** Факторните тегла на въпросите във всички факторни решения, с изключени на тези по математика и семантика, са умножени с коефициент -1.00.*

Първият въпрос, който възниква в контекста на тези наблюдения, е дали теглото на дадена променлива по съответния фактор е достатъчно високо, за да бъде тази променлива отнесена към него и взета под внимание при неговата интерпретация. Ако разглеждаме факторните тегла като коефициенти на корелация между променливите и факторите, към оценката на силата на тяхната взаимовръзка следва да се подходи по аналогичен начин. Поради това е от съществено значение определянето на някаква долна граница на „приемливост" на факторните тегла по техните абсолютни стойности, за да бъдат съответните променливи включени или извадени от процеса на интерпретация на факторите, а дори и на конструиране на съответната скала.

Тъй като няма добре разработени и надеждни тестови процедури за оценка на значимостта на факторните тегла (Kim & Mueller, 1978; Riekert & Eakin, 2008), в различни публикации авторите предлагат различни практически правила за определяне на праговите стойности. Така например де Аяла приема, че факторни тегла, по-големи от 0.50, са достатъчно високи, за да аргументират решението за отнасяне на дадена променлива към даден фактор (de Ayala, 2009). Авторът разглежда като умерено високи факторните тегла, по-големи от 0.30. Същите прагови стойности посочват и други автори, отбелязвайки, че стойности над 0.50 са желани и индикиращи стабилни, значими фактори (Darlington, 1977; Costello & Osborne, 2005). К. Рийкерт и М. Еакин отбелязват,

че стойности над 0.30 следва да се разглеждат като минимално равнище на приемливост, докато по-значими и съществени са факторни тегла над 0.40 (Riekert & Eakin, 2008). Тези автори представят и кратък преглед на публикации в областта на психологията, в които са представени резултати от прилагането на факторни анализи. В по-голяма част от тях авторите са използвали прагови стойности от 0.40 – 0.45. А. Буджа и Н. Еюбоглу, както и други автори, предлагат още по-консервативни критерии с прагови стойности от 0.50 и дори 0.80 (Buja & Eyuboglu, 1992; Costello & Osborne, 2005).

При избора на долна граница на факторните тегла е важно да се вземе предвид и връзката между факторните тегла на променливите в дадена скала и нейната надеждност - надеждността на скалата е толкова висока, колкото е квадратът на най-високото факторно тегло (Kim & Mueller, 1978).

В настоящото изследване ще приемем един умерено консервативен праг на приемливост на факторните тегла на равнище $\pm$0.45. От тази позиция структурата на факторите изглежда по различен начин. В горната таблица клетките на въпросите с факторни тегла над праговата стойност, които обозначават принадлежността им към съответния фактор, са маркирани в сиво. Лесно може да се определи, че дори и при този невисок критерий, факторните тегла със стойности над 0.45, представени като дял от всички факторни тегла (клетки) в горната таблица, съставляват едва 38%. При тестовите варианти 141 и 171 този дял е още по-нисък – 30%

След като приетият критерий за минимална стойност на факторните тегла бъде приложен, става ясно, че нито един от субтестовите фактори не може да бъде конституиран от всички въпроси, които са включени в съответния раздел. Разбира се, и в това отношение субтестовете не са равностойни. С по-голям брой айтеми с факторни тегла над праговото равнище се характеризират такива субтестове като 1. *Български език* (6 въпроса), 5. *Математика* ((7), 9. *Разсъждения* (9) и 10. *Семантика* (4). Към останалите субтестове следва да се отнесат до 3 въпроса, а като рядък феномен се очертава субтестът по география, който съдържа само един въпрос, както и този по биология - с нито един въпрос със значително, надпрагово факторно тегло.

Разбира се, броят на въпросите, които характеризират даден фактор, както и големината на техните факторни тегла, намират израз както в собствените стойности на съответния фактор, така и в неговата обяснителна сила по отношение на дисперсията. В този смисъл наблюденията върху данните от горната таблица се съгласуват напълно с тези върху таблица 5. Посочените по-горе 4 субтеста са точно тези, които се характеризират с устойчив, значим първи фактор в различните тестови варианти.

Анализите на факторните тегла на въпросите и количеството на въпросите със значими стойности отвеждат към следващия важен проблем – за постигането на прости (чисти) факторни структури в смисъла, определен Л. Л. Търстоун (Thurstone, 1934; 1935; 1936). Правилата, определени от него, са предназначени за оценка на многофакторни структури, но е очевидно, че получените еднофакторни решения представляват

възможно най-простите факторни структури. Поставянето на този въпрос обаче има смисъл поради обстоятелството, че първите поред фактори в някои субтестове, въпреки че имат по-високи собствени стойности от останалите в същия субтест, не надхвърлят критичните граници, определени на случайно равнище. Във вариант 134 с такива фактори са 2 субтеста, във вариант 141 – 6, а във вариант 171 – 3 субтеста. Като част от проблема за простите факторни структури можем да добавим сравнително ниската кумулативна обяснителна сила на факторите от първоначално извлечените конфигурации и още по-слабата обяснителна сила на финалните еднофакторни модели.

По повод постигането на проста факторна структура някои автори отбелязват, че ако даден фактор е маркиран с високи факторни тегла (от 0.50 и по-високи) по помалко от 3 до 5 променливи, този фактор е слаб и неустойчив (Kim & Mueller, 1978; Costello & Osborne, 2005). Ние ще приемем като долна граница за характеризиране на обяснителната сила, на значимостта и съдържателната „плътност" даден фактор наличието на минимум 3 въпроса с факторно тегло над 0.45, по които този фактор да бъде маркиран.

Таблица 8. Брой на въпросите със значими факторни тегла

Вариант	1. Български език	2. Литература	3. История	4. География	5. Математика	6. Физика	7. Химия	8. Биология	9. разсъждения	10. Семантика
вар. 134	6	3	3	1	7	2	3	0	9	4
вар. 141	6	1	2	2	7	2	0	2	4	5
вар. 171	4	3	4	1	7	1	3	2	3	2

Данните в таблицата показват, че като цяло субтестовите структури са нестабилни и по отношение на тази своя характеристика. С изключение на субтест 5. *Математика*, всички останали субтестови факторни структури варират по отношение на броя на въпросите със значими факторни тегла, които ги обозначават. Трябва да отбележим обаче, че в тестовите варианти, непредставени в този текст, броят на въпросите с високи факторни тегла в субтестовете по математика също варира, но в тесни граници. Устойчива съдържателна плътност демонстрират субтестове 1. *Български език*, 5. *Математика* и 9. *Разсъждения*. Към тях бихме могли да добавим и по-малко устойчивите 3. *История* и 10 *Семантика*. Останалите субтестове с номера 2. *Литература*, 4. *География*, 6. *Физика*, 7. *Химия* и 8. *Биология* се характеризират със слаби, неустойчиви фактори с ниска съдържателна плътност. Ще припомним, че именно тези субтестове се асоциират най-често с ниски, под праговото равнище собствени стой-

ности и с най-ниска обяснителна сила на еднофакторните им конфигурации.

Тези особености, без съмнение, са свързани с ниските равнища на корелация между айтемите. Възможно е, обаче, въпросите да корелират слабо само на субтестово равнище, на каквото са направени анализите в тази глава на разработката. Правдоподобно е да се направи предположението, че някои въпроси, особено тези с ниски или отрицателни факторни тегла, изпитват влиянието на друг или други фактори извън съответния субтестов фактор, което следва да се прояви в по-високите им корелации с въпроси извън субтеста, към който формално принадлежат. Това предположение ще бъде проверено по-нататък в текста.

Тук следва да отбележим и още един феномен – някои от въпросите корелират негативно със съответния фактор. Негативните корелации са малко на брой (общо 5 въпроса от различни субтестове на вариант 134) и слаби по големина, например въпрос 4 от раздела по литература с факторно тегло -0.129. Наблюдават се обаче и по-високи негативни корелации, някои от които се доближават до приетия критерий, например при въпрос 2 от раздела по история с факторно тегло -0.439. Във факторните матрици на вариант 171 негативните корелации са 9, а в тези на вариант 134 – 16. Наличието на негативно корелиращи айтеми в субтестови скали, които в тематично отношение са напълно еднородни, е феномен, на който следва да се обърне внимание.

Отрицателното факторно тегло на даден въпрос е свидетелство за това, че факторният бал по съответния субтест е свързан по противоположен начин с оценките на и. л. по дадения въпрос. С други думи, лицата, които се характеризират с високи равнища на способности по съответния субтест, по-скоро се провалят при отговора на този въпрос. Следователно, от гледна точка на философията на факторния анализ, съответният латентен фактор влияе върху вероятността от коректен отговор по противоположен, негативен начин. Тази интерпретация на негативните факторни тегла съдържа в почти чист вид определението на дискриминативната сила, както се разглежда това понятие в двете психометрични теории. Поради това би било интересно да разгледаме какви са статистиките на въпросите във вариант 134 с отрицателни факторни тегла. Индексите на тези въпроси, определени в рамките на СТТ, са следните:

Вариант 134, раздел 2. *Литература*, въпр. 4:

p = 0.23, D = 0.16, r_{bis} = -0.08 (факторно тегло -0.129)

Вариант 134, раздел 3. *История*, въпр. 2:

p = 0.04, D = 0.02, r_{bis} = -0.21 (факторно тегло -0.439)

Вариант 134, раздел 7. *Химия*, въпр. 4:

p = 0.31, D = 0.29, r_{bis} = 0.03 (факторно тегло -0.046)

Вариант 134, раздел 7. *Химия*, въпр. 7:

p = 0.63, D = 0.28, r_{bis} = -0.01 (факторно тегло -0.180)

Вариант 134, раздел 8. *Биология*, въпр. 5:

p = 0.63, D = 0.23, r_{bis} = -0.03 (факторно тегло -0.073)

Разглежданите въпроси се характеризират с ниски стойности на класическия индекс на дискриминативност и с близки до нула, в повечето случаи - негативни коефициенти на бисериална корелация, които се използват като паралелна мярка за оценка на същата характеристика. Тези наблюдения водят не само към предположението за възможно типологично сходство на дискриминативния индекс и факторните тегла на въпросите, но и към ново разбиране на неговото психометрично значение като основна характеристика на айтемите. Корелационните анализи на тази предполагаема връзка между двата вида статистики, направени върху отделните субскали от вариант 134, по-скоро я потвърждават. Коефициентите на корелация с класическия индекс D са на равнища 0.30 – 0.60, в някои случаи незначими при $\alpha = 0.05$. Значително по-високи и статистически значими са наблюдаваните взаимовръзки с бисериалния коефициент, които са на равнища 0.90 – 1.00, в повечето случаи приближаващи се до горната граница. Паралелът, който може да бъде направен между бисериалния индекс на дискриминативност и факторните тегла на въпросите е този, че индексът е оценка на взаимовръзката между даден айтем и суровия тестов бал (манифестирана променлива) така, като факторното тегло е оценка на взаимовръзката между този айтем и съответния фактор (латентна променлива).

2.3.1.6. Съотношения между първия и втория фактор

При анализа на размерността на латентните структури на равнище субтест специален интерес представлява първият фактор, съответно неговата собствена стойност (делът на обяснената от него дисперсия), и съотношението му с останалите фактори, по-специално със следващия (втори) по ред. Емпиричните собствени стойности на първия фактор при различните субтестове варират в широкия интервал от 0.556 (вариант 141, субтест 7. *Химия*) до 3.086 (вариант 171, субтест 5. *Математика*). Данните в следващата таблица дават по-ясна представа за големината на тази статистика при различните субтестове от анализираните варианти 134, 141 и 171.

Таблица 9. Групирано честотно разпределение на емпиричните собствени стойности на първия фактор

λ_{F1}	Честота	Кумулативна честота	% от всички случаи	Кумулативен % от всички случаи
0.50<x<=1.00	6	6	20.00	20.00
1.00<x<=1.50	12	18	40.00	60.00
1.50<x<=2.00	3	21	10.00	70.00
2.00<x<=2.50	7	28	23.33	93.33
2.50<x<=3.00	1	29	3.33	96.67
3.00<x<=3.50	1	30	3.33	100.00

Като цяло извлечените първи фактори се характеризират с ниски собствени стойности, като най-голямата група от 12 първи фактора (40.00% от всички) имат собствени стойности в интервала от 1.00 до 1.50. Делът от цялата дисперсия на въпросите, която може да бъде обяснена чрез първия фактор в съответния субтест, също варира в широките граници от 5.562% до 30.862% при същите два субтеста.

Таблица 10. Групирано честотно разпределение на дела от цялата дисперсия, обяснена чрез първия фактор

% от цялата дисперсия	Честота	Кумулативна честота	% от всички случаи	Кумулативен % от всички случаи
5.00<x<=10.00	6	6	20.00	20.0
10.00<x<=15.00	12	18	40.00	60.0
15.00<x<=20.00	3	21	10.00	70.0
20.00<x<=25.00	7	28	23.33	93.3
25.00<x<=30.00	1	29	3.33	96.7
30.00<x<=35.00	1	30	3.33	100.0

Данните от таблицата по необходимост следват тези от предходната, доколкото собствената стойност на фактора е мярка за дела от дисперсията, която може да бъде обяснена чрез него. От данните следва, че най-голям е броят на факторите, които обясняват 10% - 15% от цялата дисперсия. (12 фактора, 40.00% от всички субтестове). При тази ситуация, дори и да пренебрегнем резултатите от паралелния анализ, валидността предположението за еднофакторна латентна структура на субтестовете може да бъде поставена под въпрос. Приемайки едномерен модел, следва да се лишим приблизително от 70% - 95% от информацията, съдържаща се в корелационните матрици.

Важен аспект от верификацията на предположението за едномерност на субтестовите латентни структури е не само големината на собствената стойност на първия фактор, но и доколко този фактор се разграничава отчетливо от останалите фактори. Дали той може да бъде идентифициран като „главен” и „доминиращ” над останалите, които от своя страна могат да бъдат третирани като „второстепенни” и „несъществени”. И тъй като методологията на факторния анализ, както бе отбелязано, се базира на стратегията на последователно извличане на фактори с все по-намаляваща обяснителна сила, важна е съпоставката между първите два фактора от първоначално извлечените конфигурации. То може да бъде изразено чрез съотношението между техните собствени стойности ($\lambda_{F1} : \lambda_{F2}$).

Може да се каже, че, погледнати от този ъгъл, факторните структури се представят в нова, неочаквана светлина. Съотношенията между големините на собствените стойности на първия и втория фактор варират в широките граници от 7.798 при вари-

ант 134, субтест 1. *Български език* до 1.544 при вариант 141, субтест 3. *История*, т.е. доминантността на първия фактор при различните субтестове не е еднакво изразена. Данните от следващата таблица, на която е представено групираното разпределение на тези съотношения, сочат, че типичното съотношение е от 3:1 до 4:1 (при 9 субтеста, 30.00% от всички). Не е малък и делът на субтестовете с по-високи съотношения.

Нека да се върнем към предложените от Ф. Лорд два критерия за едномерност, основани на съпоставяне на собствените стойности на латентните фактори: (1) ако първият фактор е „голям в сравнения с втория" и (2) ако вторият фактор не е „много по-голям от който и да е от останалите" (Lord, 1980, стр. 21), както и на по-конкретните предложения на някои автори, съгласно които съотношението между собствените стойности на първия и втория фактор следва да е поне 3:1 (Reckase, 1979; Cooke et al., 1999; Pollard et al., 2009).

Таблица 11. Групирано честотно разпределение на съотношенията между големините на първия и втория фактор

$\lambda_{F1} : \lambda_{F2}$	честота	кумулативна честота	% от всички случаи	кумулативен % от всички случаи
1.00<x<=2.00	5	5	16.67	16.67
2.00<x<=3.00	8	13	26.67	43.33
3.00<x<=4.00	9	22	30.00	73.33
4.00<x<=5.00	6	28	20.00	93.33
5.00<x<=6.00	1	29	3.33	96.67
6.00<x<=7.00	0	29	0.00	96.67
7.00<x<=8.00	1	30	3.33	100.00
липсващи	0	30		100.00

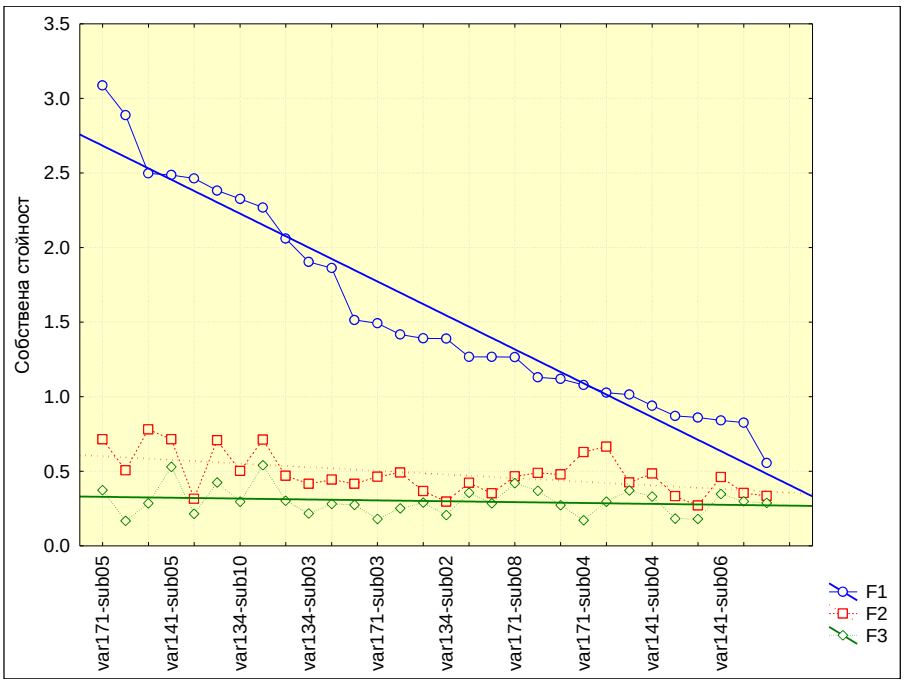
При тези условия поне 17 субтеста (56.66% от всички), при които съотношението $\lambda_{F1} : \lambda_{F2}$ е по-високо от 3, следва да се разглеждат като едномерни.

На какво се дължат големините на съотношенията $\lambda_{F1} : \lambda_{F2}$ при различните субтестове и от какво се обуславят разликите между техните големини? Следващата графика разкрива интересна особеност. На нея са представени собствените стойности на първите три фактора в анализираните 30 субтеста, подредени в низходящ ред на собствените стойности на първия по ред фактор.

Графиката илюстрира видимата разлика между собствените стойности на тези фактори в различните субтестове, чиито равнища прогресивно намаляват (λ_{F1} варира от 3.086 до 0.556). Но може би по-интересното наблюдение е това, че собствените стойности на втория по ред фактор са сравнително равностойни (λ_{F2} варира от 0.780 до 0.272) и с централна тенденция $\overline{X}$ = 0.48. Линейната апроксимация показва слаба тенденция към намаляване на техните равнища. В графиката са включени и собствените стойности на третия по ред фактор, също така сравнително равностойни, с цент-

рална тенденция $\overline{X}$ = 0.30. Линейната апроксимация показва още по-слаба тенденция към намаляване, дори за устойчивост на техните равнища.

Фигура 16. Съотношение между собствените стойности на първия, втория и третия по ред фактор при 30 субтеста



Следователно големината на съотношението между собствените стойности на първия и втория по ред фактор в отделните факторни структури, която се разглежда като белег за доминантността на първия фактор (и за едномерност на съответната структура), се дължи във висока степен на собствената стойност на първия фактор, на неговата отдалеченост от сравнително неизменния по големина втори фактор. Направеният корелационен анализ между собствените стойности на първия и втория фактор, от една страна, и техните съотношения, от друга, потвърждава подобен извод. Собствените стойности на първия фактор λ_{F1} корелират силно и позитивно с тези съотношения (r = 0.70, p < 0.05), докато корелацията на собствените стойности на втория фактор λ_{F2} е пренебрежимо слаба и очаквано негативна (r = -0.16, p > 0.05).

Направеният анализ на доминантността на първия фактор, основана на дистанцията му от следващия по ред фактор показа, че при повече от половината от анализираните субтестови факторни структури първият критерий на Ф. Лорд е удовлетворен, т .е. може да се говори за наличието на доминиращ (първи по ред) фактор. При всички субтестове е удовлетворен и следващият критерий, съгласно който вторият фактор не е много по-голям от който и да е от останалите, в случая от третия фактор.

Визуално потвърждение на тези наблюдения може да бъде намерено в графиката на фигура 13, на която са представени профилите на всички субтестове от вариант 171. По същество тя представлява диаграма (*scree plot*) на Кетел на собствените

стойности на факторите от отделните конфигурации, в която се наблюдават всички описани по-горе особености. Без съмнение, наблюдава се сходство в профилите на отделните субтестови факторни конфигурации в този вариант. Същото сходство се наблюдава, между впрочем, и при останалите тестови варианти. Профилите от графиката се характеризират с ясно изразен, стръмен преход от първия към втория фактор, последван от по-хоризонтална част, в която се намират останалите фактори (факторни отломки, по терминологията на Р. Кетел). Всичко това говори, че факторните конфигурации се характеризират с изявен първи фактор, последван от няколко значително по-слаби фактори, сред които единствено вторият при някои субтестове се отличава с по-високи собствени стойности. Същевременно се наблюдава ясно изразена диференциация между първите по ред фактори, които се характеризират с различни по големина собствени стойности. Доминантността на първия фактор, следователно, не е изразена в еднаква степен.

Ако като критерий за оценка на размерността на латентните структура се приложи графичният тест на Кетел, както препоръчват много автори (виж Hambleton & Rovinelli, 1986; Costello & Osborne, 2005 и др.), с уточнението, че броят на значимите фактори се определя според броя на собствените стойности наляво от точката на пречупване, без съмнение следва да приемем, че субтестовите факторни структури са едномерни. Тази хипотеза следва да бъде верифицирана.

2.3.1.7. Потвърдителен факторен анализ

Анализът на факторната структура на ТОП чрез анализа на главните фактори е само първият етап от изследването на латентната му структура. Бе отбелязано, че изследователският факторен анализ се разглежда като метод за разработване на психологически скали и, в по-общ смисъл, като процедура за генериране на факторни модели (Kubinger, 2003). Поради това би било уместно еднофакторният модел, които се очерта като сравнително приемлив в предходните анализи, да бъде подложен на проверка чрез използването на подходящи статистически тестове.

Едномерният модел е най-простият модел, избран и за базов в настоящото изследване. Както отбелязва Р. Дарлингтън, тъй като по-простата хипотеза логически има научен приоритет над по-сложните, хипотезите, включващи по-малък брой фактори са предпочитани пред тези, включващи повече фактори (Darlington, 1997). Но освен теоретично, налице са и множество емпирични основания. Резултатите от проведените факторни анализи сочат, че ако не всички, то поне част от субтестовите латентни структури се характеризират с наличието на един (първи по ред) фактор с по-голяма обясните сила от останалите и че, при прилагане на различни методи за определяне на броя на „значимите" фактори (паралелен анализ, критериите на Ф. Лорд или графичния тест на Кетел), тези фактори могат да бъдат квалифицирани като доминантни.

Подходящ за тази цел е потвърдителният факторен анализ (*confirmatory factor*

analysis, *CFA*), който се третира като продължение на „стандартния" изследователски факторен анализ и поради това се провежда при ясни теоретични или емпирични основи, а неговата цел е верифицирането на конкретен факторен модел. Чрез него могат да бъдат тествани специфични хипотези за размерността на факторното пространство, структурата на факторните тегла на променливите и корелациите между тях (Hurley et al., 1997; Stevens, 2002).

Като статистически метод потвърдителният факторен анализ е част от по-широката мрежа от многомерни статистически техники, познати като моделиране със структурни уравнения (*structural equation modeling*), латентно-структурно моделиране или латентно-структурен анализ. Основната идея, която стои в основата на тази група от анализи е, че чрез система от линейни уравнения могат да бъдат тествани хипотези за взаимовръзките между дадена съвкупност от променливи, като се изследват техните вариации и ковариации.

Дизайн на изследването

Целта изследването е да се верифицира предположението за наличието на еднофакторна латентна структура при субтестовете на ТОП, която обуславя отговорите на и. л. За да направим това, ще се насочим към два от субтестовете, които представляват специален интерес: (1) субтестът с най-високо съотношение между собствените стойности на първия и втория фактор и (2) субтестът с най-висока собствена стойност на първия фактор. И в двата случая, съгласно концепциите на Ф. Лорд, Р. Хъмбълтън и други автори, можем да предположим, че първият фактор е достатъчно ярко изразен, за да бъде идентифициран като доминиращ и който да обоснове допускането за едномерност на латентната структура на съответния субтест.

Предходните факторни анализи откроиха като латентна субтестова структура с най-високо съотношение между първите два фактора тази на субтест 1. *Български език* от вариант 134 със собствени стойности съответно 2.464 и 0.316 и съотношение между тях 7.798. Субтестът с най-висока собствена стойност на първия фактор е 5. *Математика* от вариант 171, със собствени стойности на първите два фактора съответно 3.086 и 0.715 и съотношение между тях 4.319. Впрочем, и двата субтеста имат латентни структури, при които съотношението между първите два фактора надхвърля 3:1 и дори 4:1, което се разглежда от някои автори като достатъчно за вземане на решение за едномерност (Reckase, 1979; Cooke et al., 1999; Pollard et al., 2009).

Определяне на променливите величини

В изследването са включени следните видове променливи

(1) зависими (манифестирани ендогенни) променливи – отговорите на и. л. по 10-те въпроса, включени в съответния субтест;

(2) независима (латентна екзогенна) променлива – един общ фактор, за които

допускаме, че влияе върху всяка манифестирана променлива;

(3) независими (латентни екзогенни) променливи – остатъчни (уникални) фактори, всеки от които оказва влияние върху отделна манифестирана променлива и които пораждат техните уникални дисперсии.

Хипотеза

Нулевата хипотеза, която ще бъде подложена на проверка е, че субтестовите латентни модели са едномерни. С други думи, наблюдаваната обща (споделена) дисперсия на тестовите въпроси в рамките на един субтест може да бъде обяснена с влиянието на един единствен фактор.

Изходни данни

Това са матриците на тетрахоричните коефициенти на корелация, изчислени по данните от всеки субтест, използвани и при изследователския етап на факторния анализ. Този тип данни са предпочетени, тъй като корелациите са по-подходящи в случаите, когато целта на анализа е моделиране на латентната структура, а не обясняване на общата вариация.

Метод за оценка на параметрите на уравненията в структурния модел

Като метод за оценка на параметрите на модела (функция на несъответствията) е използвана двустъпкова процедура GLS→ML, започваща с генерализирания метод на най-малките квадрати (*Generalized least squares*), последван от метода на максималното правдоподобие (*Maximum likelihood estimation*).

За проверка на хипотезата за едномерност е използван модулът SEPATH от статистическия пакет STATISTICA. Тестваният модел се проверява посредством специфичен програмен език PATH1, чрез който по-голяма част от моделите могат да бъдат представени чрез верижни диаграми (*path diagrams*), които играят фундаментална роля в структурното моделиране, показвайки променливите и каузалните връзки между тях.

В следващата таблица са представени резултатите от оценката на едномерния модел на субтеста по български език. Всеки ред представя връзката във верижната диаграма между съответната двойка латентна екзогенна променлива – манифестирана променлива (тестов въпрос), заедно с оценката на съответния свободен параметър и придружаващите я статистики, направена при шестата итерация при процедура със зададени като максимален брой 50 итерации.

В първите 10 реда на горната таблица са представени свободните параметри (числови коефициенти), представящи факторните тегла на манифестираните променливи по общия фактор. Успоредно с това в таблицата са представени и стандартните грешки на оценките, както и T-статистиката заедно с нейните нива на значимост.

Таблица 12. Оценки на едномерния модел

Връзки	Оценка на параметъра	Стандарт-на грешка	Т-статистика	Ниво на значимост
(Бълг. език)-1->[въпрос 1]	0.512	0.033	15.499	0.000
(Бълг. език)-2->[въпрос 2]	0.457	0.035	13.114	0.000
(Бълг. език)-3->[въпрос 3]	0.651	0.028	23.386	0.000
(Бълг. език)-4->[въпрос 4]	0.577	0.031	18.827	0.000
(Бълг. език)-5->[въпрос 5]	0.320	0.039	8.315	0.000
(Бълг. език)-6->[въпрос 6]	0.677	0.027	25.253	0.000
(Бълг. език)-7->[въпрос 7]	0.372	0.037	9.972	0.000
(Бълг. език)-8->[въпрос 8]	0.601	0.030	20.204	0.000
(Бълг. език)-9->[въпрос 9]	0.311	0.039	8.032	0.000
(Бълг. език)-10->[въпрос 10]	0.322	0.038	8.356	0.000
(DELTA1)-->[въпрос 1]				
(DELTA2)-->[въпрос 2]				
(DELTA3)-->[въпрос 3]				
(DELTA4)-->[въпрос 4]				
(DELTA5)-->[въпрос 5]				
(DELTA6)-->[въпрос 6]				
(DELTA7)-->[въпрос 7]				
(DELTA8)-->[въпрос 8]				
(DELTA9)-->[въпрос 9]				
(DELTA10)-->[въпрос 10]				
(DELTA1)-11-(DELTA1)	0.738	0.034	21.863	0.000
(DELTA2)-12-(DELTA2)	0.792	0.032	24.893	0.000
(DELTA3)-13-(DELTA3)	0.577	0.036	15.919	0.000
(DELTA4)-14-(DELTA4)	0.667	0.035	18.860	0.000
(DELTA5)-15-(DELTA5)	0.897	0.025	36.378	0.000
(DELTA6)-16-(DELTA6)	0.542	0.036	14.918	0.000
(DELTA7)-17-(DELTA7)	0.862	0.028	31.120	0.000
(DELTA8)-18-(DELTA8)	0.639	0.036	17.871	0.000
(DELTA9)-19-(DELTA9)	0.903	0.024	37.503	0.000
(DELTA10)-20-(DELTA10)	0.897	0.025	36.224	0.000

Тази статистика, която е важен показател за значимостта на съответната връзка, е асимптотична нормална статистика, близка до t-теста на Стюдънт, без да има същото разпределение, и се изчислява като отношение между оценката на параметъра и стандартната грешка. T-статистиката е тест за проверка на нулевата хипотеза, че стойността на съответния параметър е равна на нула.

Следващите 10 реда представят остатъчните фактори (обозначени с Delta), които са свързани с някакъв специфичен, уникален аспект от съответната предметна област, който характеризира даден конкретен тестов въпрос и нито един от останалите. Последните 10 реда представят връзки, представящи факторните тегла на манифестираните променливи по уникалните фактори.

Няма съмнение, че факторните тегла на въпросите от раздела по български

език по общия фактор са относително високи, а стандартните грешки – пренебрежимо малки в сравнение с тях. Както би могло да се очаква, оценките на параметрите от горната таблица са много близки до стойностите на факторните тегла, определени чрез изследователския факторен анализ (представени в таблица 7). Нивата на статистическа значимост на проверяващата T-статистика позволяват последователното отхвърляне на хипотезите за нулеви коефициенти, което означава съхраняването им в модела. Смущение обаче предизвикват високите факторни тегла на въпросите по уникалните фактори, които системно надвишават тези, оценени за общия фактор. Това означава, че по-голяма част от цялата дисперсия се дължи на въздействието на уникалните, а не на общия фактор. Това е първият сигнал, че проверяваният едномерен модел не е адекватен на реалните данни. На следващата таблица са представени някои от основните статистики, свързани с неговата оценка.

Таблица 13. Основни статистики при оценката на модела

Статистики	Стойност
Функция на несъответствията	0.289
MRC	0.000
Критерий ICSF	0.000
Критерий ICS	0.000
RMS стандартизиран остатък	0.052
ML χ^2	205.752
Степени на свобода	35.000
Ниво на значимост	0.000

Някои от тези статистики са благоприятни за оценявания модел. Такива са *MRC* (*Maximum residual cosine*) с нулева стойност, която показва, че итеративната процедура на оценка на параметрите на модела е завършена успешно; други два критерия също с нулеви стойности като *ICSF* (*Invariance under a constant scale factor*), който указва инвариантността на модела по отношение на константен скалиращ фактор и *ICS (Invariance under change of scale)*, свързан с инвариантността на модела по отношение на промени в скалата.

Друга, по важна част от тях обаче поставят под съмнение неговата адекватност. На първо място следва да посочим стойността на функцията на несъответствията (*discrepancy function*). Това са семейство функции, чрез които се проверява до каква степен тествания модел се съгласува с емпиричните данни. Оценките на параметрите на модела се подбират така, че числовата стойност на функцията да бъде възможно най-малка. Тази стойност е винаги по-голяма от или равна на нула и нулевата стойност на функцията означава пълна адекватност на тествания модел. Следователно всяка стойност, различна от нула, означава определена степен на несъответствие, на „отдаленост” на модела от реалните данни. Наблюдаваната стойност на функцията от

0.289 е свидетелство, че тестваният едномерен модел не е най-доброто описание на латентна структура на този субтест.

Над критичната стойност от 0.05 е и наблюдаваната стойност 0.052 на критерия *RMS* (*Root mean square*) стандартизиран остатък, която е неприемлива от гледна точка на годността на модела. Основен при проверката на нулевата хипотеза е тестът за обща адекватност на проверявания модел χ^2, който се изчислява за почти всички видове функции на несъответствията. Данните в горната таблица показват изключително висока стойност на тази статистика (χ^2 = 205.752, df = 35), както и ниското ниво на асоцииранатата с нея вероятност (p = 0.000). Въз основа на резултатите от теста, хипотезата за наличието на един общ фактор, който генерира споделената дисперсия на отговорите от субтеста по български език, може да бъде отхвърлена при ниво на значимост α = 0.05.

Друга група от показатели за годността на тествания модел са класическите едноизвадкови критерии, резултатите от които са представени на следващата таблица.

Първите два теста за годност – *GFI* (*Goodness-of-fit index*) и изравнения индекс *AGFI* (*Adjusted goodness-of-fit-index*), предложени от К. Йореског и Д. Сьорбом, показват несъответствие между тествания едномерен модел и реалните данни, макар че стойността на първия се приближава към критерийната стойност от 0.95.

Таблица 14. Едноизвадкови индекси за адекватност на модела

Тест за адекватност	Наблюдавана стойност	Приемлива стойност
GFI на Йореског	0.948	>0.95
AGFI на Йореског	0.918	>0.95
Нормиран индекс за годност на Бентлер-Боне	0.835	$\cong$ 1.00

По-значимо е несъответствието, фиксирано от следващия тест на Бентлер-Боне (*Bentler-Bonett normed fit index*), чиято стойност клони към 1.00 при пълна адекватност на модела. Като цяло, едноизвадкови индекси свидетелстват за неадекватността на тествания едномерен модел (критичните стойности на индексите, показващи висока степен на съгласуваност, са по Steiger, 2009).

Трета група от показатели, които все по-често се използват за оценка на общото съответствие на тествания модел с данните, изместващи хи-квадрат и едноизвадковите критерии, са базирани на оценката на популационния нецентрален параметър (*Population noncentrality parameter*). Концепцията е разработена от Дж. Стайгър и Дж. Линд (Steiger & Lind, 1980; Steiger, Shapiro & Browne, 1985; Steiger, 1994) поради това, че, според Дж. Стайгър, „класическият подход на тестване на хипотези е неподходящ […] поради недостатъчна мощност на теста хи-квадрат” (Steiger, 2009, стр. 1). С тази концепция авторите правят коренна промяна в подхода за оценка на адекватността на

тествания модел. Вместо проверка на нулевата хипотеза за пълна адекватност на модела, те предлагат противоположния подход – да се търси до каква степен е неадекватен, колко отдалечен е моделът от генералната съвкупност и доколко точно е определена тази неадекватност на базата на извадковите данни.

Таблица 15. Стойности на критериите, базирани на популационния нецентрален параметър

Критерий	90% долна граница на дов. интервал	Точкова оценка	90% горна граница на дов. интервал	Приемлива стойност
Популационен нецентрален параметър	0.171	0.228	0.295	
RMSEA индекс на Стайгер-Линд	0.070	0.081	0.092	<0.05
Нецентрален индекс на МакДоналд	0.863	0.892	0.918	>0.95
Популационен индекс гама	0.944	0.956	0.967	>0.95
Изравнен популационен индекс гама	0.912	0.932	0.948	>0.95

Индексите, базирани на оценка на популационния нецентрален параметър, освен стандартните точкови оценки, позволяват пресмятане и на техните доверителни интервали. На горната таблица са представени някои от най-важните статистики, базирани на този параметър, с техните точкови оценки и 90% доверителни интервали.

Критерият на Стайгер и Линд *RMSEA* (R*oot mean square error of approximation*) е основан пряко на популационния нецентрален индекс и се определя по формулата:

$$R^* = \sqrt{\frac{F^*}{\nu}} \qquad (62)$$

където:

F^* - оценка на популационен нецентрален индекс

ν - степени на свобода

Този критерий преодолява един от недостатъците на едноизвадковите методи, компенсирайки простотата (*parsimony*) на модела. При равни други условия, един модел с по-малко параметри се съгласува по-слабо от един по-сложен модел. Поради това индексите за годност, които не отчитат това обстоятелство, могат по-често да доведат до отхвърляне на тествания модел. В случая нито точковата оценка на горното съотношение (0.081), нито стойностите между двете граници на доверителния интервал дават основание да се приеме, че тествания едномерен модел е адекватен на данните. Индексът на МакДоналд *(McDonald noncentrality index*) не компенсира простотата на модела и поради това би могъл да бъде по-неточен от предходния. Независимо от това, тъй като тестваният едномерен модел не е опростен и включва всички манифестирани променливи, този индекс може да бъде използван. Неговите стойности

също свидетелстват за липса на съгласуваност между тествания модел и данните. Точковата оценка на индекса (0.892), както и интервалните му оценки, са далеч под критичната му стойност от 0.95. Популационният индекс гама (*Population gamma index*), предложен от Дж. Стайгър, И. Танака и Г. Хуба, отчита, освен популационния нецентрален индекс, и броя на манифестираните променливи по следната формула.

$$\Gamma_1 = \frac{p}{2F^* + p} \qquad (63)$$

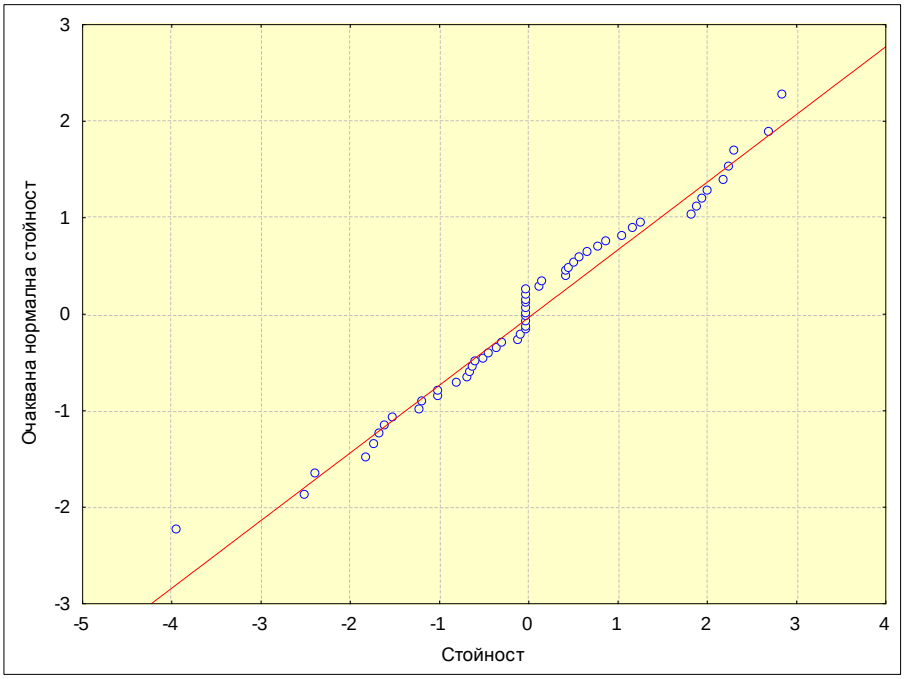
където:

F^* - оценка на популационния нецентрален индекс

p – брой на манифестираните променливи

Този индекс е разширение на индекса *GFI* на К. Йореског и Д. Сьорбом, който представлява извадкова статистика, оценяваща популационния *GFI*. Единствено този критерий дава индикации за годност на тествания едномерен модел, тъй като точковата му оценка (0.956) и част от интервалните му оценки са в полето на приемливите стойности. Изравненият популационен индекс гама (*Adjusted population gamma index*) е аналогичен на *AGFI* на К. Йореског и Д. Сьорбом, но компенсира простотата на модела. Неговите стойности – като точкова и интервални оценки, също не дават основание да се приеме, че тествания едномерен модел на субтеста по български език е адекватен на данните. Още едно свидетелство, този път графично, за адекватността на тествания едномерен модел е нормална вероятностна графика на нормализираните остатъчни стойности, представена на следващата фигура.

Фигура 17. Нормална вероятностна графика на остатъчните стойности в корелационната матрица на субтеста по български език, вариант 131



Остатъчните стойности представят разликите между емпиричните данни в корелационната матрица и възпроизведената корелационна матрица. Очевидно е, че при по-висока степен на съгласуване на тестовия модел с данните, остатъците ще бъдат по-малки. При пълно съгласуване се очаква те да бъдат нормално разпределени. Формата на кривата наподобява двойно S и не се вмества дори в интерпретативните схеми, предложени от Тюки и съструдници (Hoaglin, Mosteller & Tukey, 1991, стр. 187). Но без съмнение тя свидетелства за отклонение на формата на разпределението на остатъчните стойности от нормалното. Това, от своя страна, е свидетелство за недоброто съгласуване на тествания едномерен модел с реалните данни.

Резултатите от проверката на хипотезата за едномерност на субтест 5. *Математика* от вариант 171 не се различават съществено от тези, представени по-горе. Ще припомним, че това е субтестът с най-висока собствена стойност на първия фактор (3.086) и със съотношение между собствени стойности на първите два фактора 4.319.

Таблица 16. Резултати от потвърдителния факторен анализ върху данни от субтест 5. *Математика*, вариант 171

		Статистики			**Приемливи стойности**
Изходни данни		тетрахорични корелации			
Метод за оценка		GLS-ML			
Стойност на функцията на несъответствията		0.963			
RMS стандартизиран остатък		0.092			
χ^2 (df)		798.120 (35)			
равнище на значимост		0.000			
едноизвадкови индекси	GFI на Йореског	0.841			>0.95
	AGFI на Йореског	0.750			>0.95
	Индекс на Бентлер-Боне	0.699			≅ 1.00
критерии, базирани на популационния нецентрален параметър		90% д. гр. на дов. интервал	Точкова оценка	90% г. гр. на дов. интервал	Прагова стойност
	Популационен нецентрален параметър	0.799	0.904	1.019	
	RMSEA индекс на Стайгер-Линд	0.151	0.161	0.171	<0.05
	Нецентрален индекс на МакДоналд	0.601	0.636	0.671	>0.95
	Популационен индекс гама	0.831	0.847	0.862	>0.95
	Изравнен популационен индекс гама	0.734	0.759	0.784	>0.95

В горната таблица са представени най-важните статистически резултати от потвърдителния факторен анализ. Ще обърнем внимание на високата стойност на функцията на несъответствията (0.963), както и на нивото на статистическа значимост на теста χ^2 (p = 0.00), което дава основание за отхвърляне на нулевата хипотеза за пълна адекватност на тествания едномерен модел при ниво на значимост α = 0.05.

В съгласие с този резултат са и данните от приложените едноизвадкови тестове на Йореског и Бентлер-Боне, а също така и от поредицата тестове, базирани на популационния нецентрален параметър. Всички те свидетелстват за неадекватност на тествания модел.

Нормална вероятностна графика на остатъчните стойности от корелационната матрица на субтеста по математика, вариант 171 е още една графична илюстрация в подкрепа на тези наблюдения. Формата на кривата е още по-сложна от предходната - наподобява тройно S и също подсказва за отклонение на формата на разпределението на остатъчните стойности от нормалното. Това, от своя страна, е свидетелство за недоброто съгласуване на тествания едномерен модел на субтеста по математика с данните от корелационната матрица.

Случай 2. Факторна структура на ТОП на равнище цялостен тест (компоненти на анализа са всички въпроси в теста, к=100)

2.3.2.1. Извличане на първоначалните (незавъртени) факторни конфигурации

Разглеждането на ТОП като цялостен тест, като единна "суперскала", чиито компоненти не са субтестовете, а отделните въпроси, на пръв поглед е необосновано. От гледна точка на неговия дизайн, ТОП е замислен като тестова батерия със самостоятелни, независими един от друг субтестове, представящи различни, обособени дисциплини. Такъв подход би означавал, по метафоричния израз на Дж. Глас и Дж. Стенли, да се смесват ябълки и круши в една кошница. От друга страна, както остроумно забелязват П. Суппес и Дж. Зинес, „…когато обучаваме някого в основите на науката, ние постоянно го предупреждаваме, че „няма смисъл" […] да се събират числа, които се отнасят към различни свойства, да кажем тегло и ръст, но в същото време, без да се замисляме, предлагаме на изучаващия физика да умножава числа, свързани с такива понятия като скорост и време […]. Защо умножението е „по-смислено" от събирането?" (Суппес и Зинес, 1967, стр. 9).

Анализът на равнище цялостен тест е необходим поради две основни причини. Първата от тях е прагматична и се изразява в това, че общият тестов бал се използва като балообразуващ компонент в процедурата за подбор на студенти в някои бакалавърски програми в Нов български университет. С други думи, тестът се разглежда като единен инструмент, който отразява глобално равнището на общообразователна подготовка на кандидатите. След като общият тестов бал играе основна роля за опреде-

ляне на постиженията на кандидатите, анализът на неговата факторна структура като част от изследването на приложимостта на психометричните теории е от съществено значение. Втората причина е съдържателна и поради това – далеч по важна. Тя произтича от множеството находки в предходния анализ на субтестово равнище, които подсказват за наличието на корелационни връзки между айтемите, простиращи се отвъд формалните субтестови рамки. Ако това предположение намери основание в реалните данни, то би могло да се окаже, че латентната структура на ТОП организирана на принцип, различен от вложения в него тематичен принцип.

При анализа на латентната структура на ТОП на равнище цялостен тест бе приложена същата изследователска методология, която бе следвана и при анализа на субтестово равнище. Схематично тя може да бъде представен по следния начин: (1) генериране на корелационни матрици на тетрахоричните коефициенти на корелация между въпросите, (2) извличане на първоначалните (незавъртени) факторни конфигурации по метода анализ на главни фактори, субметод на главните оси (*Principal axis factoring*), (3) определяне на факторните модели (определяне на оптималния брой на факторите) по метода на паралелния анализ, (4) анализ на факторните тегла на въпросите, (5) анализ на съотношенията между собствените стойности на факторите и (6) потвърдителен факторен анализ.

Тъй като в предходната част на разработката бе представен не само изследователският подход, но и всички мерки, предприети за осигуряване на вътрешната валидност на изследването, по-точно на неговата статистическа и екологична валидност (Cook & Campbell, 1979; Cohen & Swerdlik, 2005), тук ще бъдат очертани само по-важните моменти.

Процедурите за извличане на факторите от отделните корелационни матрици по метода на главните оси бяха извършени при такива първоначални конфигурации на факторните модели, които предполагат наличието на 100-факторни структура, т.е. при допускане, че броят на латентните фактори във всеки тест е равен на броя на зависимите променливи (тестови въпроси), при зададена минимална собствена стойност на факторите $\lambda_{Fi} = 0.00$. Прилагането на тази процедура обаче бе възпрепятствано, при почти всички анализирани тестови корелационни матрици, от наличието на (мулти)колинеарност между променливите – особеност, която бе наблюдавано и на субтестово равнище.

За преодоляване на този проблем бе приложена процедурата на М. Пет и сътрудници (Pett, Lackey & Sullivan, 2003) за отстраняване на онези променливи, които корелират високо с една или повече други променливи.

При вариант 134 от анализа бяха отстранени три въпроса с номера 33 (поради висока корелация с въпрос 34, $r_{tet} = 1.000$), 42 (висока корелация с въпрос 45, $r_{tet} = 0.572$) и 81 (висока корелация с въпрос 84, $r_{tet} = 0.545$). При вариант 141 от анализа бяха извадени шест въпроса с номера 1 (поради висока корелация с въпрос 26,

r_{tet} = 0.607 и с въпрос 99, r_{tet} = 0.545), 2 (висока корелация с въпрос 5, r_{tet} = 0.477), 47 (висока корелация с въпрос 46, r_{tet} = 0.533), 48 (висока корелация с въпрос 67, r_{tet} = 0.477 и с въпрос 75, r_{tet} = 0.560), 90 (висока корелация с въпрос 84, r_{tet} = 0.538) и 98 (висока корелация с въпрос 27, r_{tet} = 0.500 и с въпрос 99, r_{tet} = 0.561). При вариант 171 от анализа бе изключен само един въпрос с пореден номер 41 (поради висока корелация с въпрос 48, r_{tet} = 0.686). След отстраняване на съответните въпроси факторните анализи са направени с редуцираните стартови конфигурации, в които се предвиждат съответно 97 фактора за вариант 134, 94 за вариант 141 и 99 фактора за вариант 171. Паралелните анализи със симулирани данни са направени при същите параметри. Пълните резултати от направените факторни анализи при избраната стартова конфигурация, както и резултатите от последващите допълнителни анализи, са представени в таблица в приложение 8. В следващата таблица е представено извлечение на някои по-важни данни от това приложение.

Таблица 17. Резултати от факторния анализ на главните оси на варианти 134, 141 и 171

Източник на данни	Статистики на реалните данни				Статистики на симулираните данни	
	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни ст-ти	95-ти процентил
1	2	3	4	5	6	7
вариант 134	F1	12.278	12.658	12.658	1.826	1.880
	-	-	-	-	-	-
	F10	**1.576**	**1.625**	**31.517**	**1.544**	**1.566**
	F11	1.482	1.528	33.044	1.522	1.545
	-	-	-	-	-	-
	F55	0.015	0.015	59.555	0.887	0.899
вариант 141	F1	8.071	8.586	8.586	1.801	1.854
	-	-	-	-	-	-
	F10	**1.567**	**1.667**	**27.255**	**1.522**	**1.546**
	F11	1.404	1.494	28.749	1.500	1.521
	F51	0.000	0.000	53.855	0.917	0.930
вариант 171	F1	8.601	8.688	8.688	1.767	1.818
	-	-	-	-	-	-
	F8	**1.626**	**1.642**	**22.726**	**1.554**	**1.575**
	F9	1.546	1.561	24.287	1.532	1.554
	-	-	-	-	-	-
	F55	0.010	0.011	52.004	0.909	0.920

Таблицата съдържа информация за първия и последния фактор от незавърте-

ните факторни решения, както и за факторите, след които собствените стойности са по-ниски от референтните, получени по метода на паралелния анализ. Резултатите от факторния анализ сочат, че незавъртените решения съдържат изключително голям брой фактори – 55 при вариант 134, 51 - при вариант 141 и също 55 - при вариант 171.

Подобно на първоначалните конфигурации на субтестово равнище, броят им възлиза на малко над половината от броя на въпросите. Въпреки че съдържат голям брой фактори, извлечените латентни структури обясняват малко над половината от цялата дисперсия на манифестираните променливи – между 52.00% при вариант 171 и 59.56% при вариант 134. Това е, разбира се, делът на общата (споделена) дисперсия на въпросите, която е генерирана от тази широка структура от общи фактори. Останалият, почти равностоен на нея дял от дисперсията, се дължи на уникалните за всеки въпрос или на случайните фактори.

Относително невисоката кумулативна обяснителна сила на извлечената факторна конфигурация се дължи на слабата обяснителна сила на отделните фактори. Данните от приложение 8 действително показват, че преобладаващата част от факторите се характеризират с ниски собствени стойности и следователно – с ниска обяснителна сила. По-конкретно, около 65% - 66% от факторите имат собствени стойности, по-ниски от 1.00 и съответно обясняват по-малко от 1% от дисперсията. От друга страна, големият брой на факторите в незавъртяното решение означава, че първите по ред фактори, които неизменно се характеризират с по-високи факторни тегла, не са успели да „извлекат" възможно най-голяма част от споделената дисперсия и сред множеството от въпроси все още има такива, чиято дисперсия следва да бъде обяснена с нови и нови фактори. Действително, ортогоналната ротация на суровите факторни структури по метода *varimax normalized*, още преди определяне на финалните факторните модели, показва, че почти всички фактори са маркирани с високо факторно тегло поне по един, по често – по няколко въпроса, които при някои фактори принадлежат на различни тематично обособени субтестове. Това следва да се разглежда като свидетелство, че всеки отделен фактор оказва, по-силно или по-слабо, влияние върху отделна група от сравнително малък брой въпроси, които корелират помежду си, независимо от тематичната си определеност.

2.3.2.2. Особености на взаимовръзките между въпросите

Наблюдаваният феномен – латентни тестови структури с голям брой независими фактори, може да бъде обвързан с още един феномен, който бе обсъден по-горе в текста - наличието на мултиколинеарност в корелационните матрици, което прави невъзможно прилагането на факторния анализ върху пълния им обем. Тези наблюдения повдигат един друг интересен въпрос, който има директно отношение към обсъжданата тема – за интракорелациите между въпросите от отделния субтест (предметна област) и за интеркорелациите между въпросите от различни субтестове. Дизайнът на

ТОП предполага да очакваме високи равнища на вътрешните корелации между въпросите в рамките на отделните субтестове и ниски равнища на корелацията им с въпросите от другите субтестове. Данните обаче очертават една доста по-различна картина.

Нека да се обърнем към матриците с тетрахоричните корелации на анализираните три тестови варианта и да разгледаме само коефициентите на корелация r_{tet} с абсолютна стойност, по-висока от 0.50 (т.е. с умерена и висока положителна или отрицателна корелация). При вариант 134, от общо 12 двойки въпроси с корелационни коефициенти в очертания интервал, само в 4 двойки (33.33%) въпросите принадлежат на една и съща предметна област, а в останалите 8 двойки (66.66%) – на различни предметни области. При вариант 141, от общо 10 двойки въпроси със същите равнища на корелации, само в 3 двойки (30.00%) въпросите принадлежат на един и същи субтест, а в останалите 7 двойки (70.00%) – на различни предметни области. И накрая, при вариант 171, от общо 8 двойки въпроси с корелации над 0.50, в 4 двойки (50.00%) въпросите принадлежат на един и същи субтест, а в останалите 4 двойки (50.00%) – на различни предметни области. С други думи, въпросите с високи корелационни взаимовръзки, за които бихме очаквали, че ще бъдат манифестация на силни латентни променливи, по-често принадлежат на различни субтестове, отколкото на един и същи субтест. По-детайлна информация по този въпрос е представена в приложение 9.

Картината при последния разгледан тестов вариант 171 от същото приложение е интересна с това, че в по-голяма част от корелационните двойки въпроси участва един и същи въпрос - 41 от раздел 5. *Математика*. Той се характеризира с високи корелации в разглеждания интервал не само с въпроси от същата предметна област, но и с такива от други области, за които лесно би могло да се предположи (напр. химия) или, обратно, трудно би могло да се предположи наличие на взаимовръзки (семантика, разсъждения). Същевременно този въпрос има далеч по-слаби взаимовръзки с някои други въпроси от същия субтест, например практически нулева корелация с въпрос 50 (r_{tet} = 0.010). Би било трудно да си представим, че отговорите на и. л. на тези два математически въпроса са повлияни от един и същи латентен фактор.

Такива съюзи между тематично разнородни въпроси не са рядкост и в други тестови варианти. Например при вариант 134 участник в такава странна група е въпрос 81 от раздел 9. *Разсъждения*, който корелира високо с въпрос 84 от същия раздел (r_{tet} = 0.545), но и с въпроси 22 от раздел 3. *История* ((r_{tet} = -0.507) и 58 от раздел 6. *Физика* (r_{tet} = 0.529). Във вариант 141 въпрос 48 от раздел 5. *Математика* е свързан по подобен начин с въпрос 38 от раздел 4. *География* (r_{tet} = -0.653), с 54 от раздел 6. *Физика* (r_{tet} = -0.659), с 59 от същия раздел (r_{tet} = -1.000) и с въпрос 75 от раздел 8. *Биология* (r_{tet} = 0.560).

Става все по-ясно, че корелационните валенции на въпросите се простират отвъд рамките на предметните области, зададени чрез съответните субтестове. Забелязва се и друга особеност, че някои въпроси образуват около себе си мрежа от силни

корелационни връзки с въпроси, които принадлежат на същата или на други предметни области. Тези въпроси могат да бъдат характеризирани като „корелационни възли", а пример за такъв корелационен възел е въпрос 41 от субтеста по математика от вариант 171, който има валенции към въпроси във и извън предметната област, към която принадлежи.

Интересно би било да се установи дали интракорелациите между въпросите в даден субтест са по-високи от интеркорелациите им с въпросите от останалите субтестове, както бихме могли да очакваме. Подходяща мярка за оценка е коефициентът на множествена корелация (*multiple R*), който генерализира коефициента на корелация, и неговата производна – коефициентът на детерминация (*multiple* R^2). Двете мерки се използват при метода на множествената регресия за оценка на качеството на изградения модел, но могат да бъдат полезни при оценката на интра- и интеркорелациите, ако приемем (условно) даден въпрос като зависима променлива, а останалите въпроси от субтеста (от целия тест) – като предиктори. R^2 се използва и във факторния анализ като базов метод за оценка на дисперсията, която всяка променлива споделя с всички останали (*communalities*). Тази мярка е интересна с това, че описва обяснената дисперсия и подлежи на пряка интерпретация. Този коефициент, който представлява отношението между обяснената дисперсия и цялата дисперсия на дадена променлива, показва каква част от вариацията на зависимата променлива y се дължи на вариацията на стойностите в предикторите x_1, x_2, x_3, …, x_n. Подобно на коефициентите на единичната и на множествената корелация, той отразява силата на линейна връзка между променливите и може да приема стойности в интервала $0.00 \leq R^2 \leq 1.00$. Умножен по 100, изразява силата на влияние на независимите променливи в проценти.

Таблица 18. Интра- и интеркорелации между въпросите от вариант 171, субтест 5. *Математика*

Въпроси	Интракорелации		Интеркорелации	
	коефициент на множествена корелация *R*	коефициент на детерминация R^2	коефициент на множествена корелация *R*	коефициент на детерминация R^2
41	*0.277	*0.077	*0.395	*0.156
42	*0.282	*0.080	*0.422	*0.178
43	*0.316	*0.100	*0.389	*0.152
44	*0.275	*0.075	*0.427	*0.182
45	*0.336	*0.113	*0.430	*0.185
46	*0.325	*0.106	*0.401	*0.161
47	*0.276	*0.076	*0.395	*0.156
48	*0.412	*0.170	*0.440	*0.193
49	*0.399	*0.159	*0.398	*0.158
50	*0.230	*0.052	*0.373	*0.139

Заб. Всички коефициенти, маркирани със звездичка (*), са значими при $\alpha = 0.05$

Като пример в таблицата по-горе са представени интракорелациите между въпросите от субтест 5. *Математика* от вариант 171 - чрез коефициентите на множествена корелация R и на детерминация R^2 на съответния въпрос с всички останали въпроси от субтеста, както и техните интеркорелации - чрез същите два коефициента на корелация на съответния въпрос с всички останали въпроси извън субтеста.

Данните от горната таблица като че ли поднасят поредната изненада. Коефициентите на множествена корелация, съответно на детерминация, отразяващи взаимовръзките между променливите в рамките на субтеста, са системно по-ниски от тези, изразяващи връзките им с въпросите от други предметни области. Не може да не бъде отбелязано обстоятелството, че въпроси като тези с номера 44 и 50 споделят обща вариация с останалите въпроси по математика в размер на 5% - 8%, докато с всички въпроси от други предметни области – 14% - 18%.

Тези данни насочват нашето внимание освен към въпроса за наличието на (силни) взаимовръзки между въпросите от различни предметни области, и към въпрос, който на пръв поглед не изглежда дискусионен - към характера на вътрешните взаимовръзки в рамките на съответния субтест.

Особено интересен, макар и не съвсем типичен, е случаят със субтест 7. *Химия* от вариант 141. От общо 45 двойки въпроси (коефициента на корелация) 21 (46.66%) свидетелстват за обратнопропорционални връзки между въпросите. Действително, част от тях са много слаби, практически нулеви, но като цяло средното равнище на негативните корелации (-0.086) е съпоставимо със средното равнище на позитивните корелации (+0.060).

Таблица 19. Интракорелации между въпросите от вариант 141, субтест 7. *Химия*

въпроси	**61.**	**62.**	**63.**	**64.**	**65.**	**66.**	**67.**	**68.**	**69.**	**70.**
61.	-									
62.	-0.004	-								
63.	0.001	-0.002	-							
64.	-0.071	-0.022	-0.127	-						
65.	-0.128	-0.146	-0.108	-0.104	-					
66.	0.105	0.017	0.148	0.142	0.040	-				
67.	0.017	**0.148**	-0.113	0.097	**-0.232**	0.022	-			
68.	-0.014	-0.134	0.064	0.015	0.027	-0.126	0.062	-		
69.	-0.093	0.031	0.042	0.097	-0.086	0.020	0.039	0.084	-	
70.	-0.015	0.047	0.108	0.006	0.061	-0.154	-0.077	-0.038	-0.011	-

Клетките в горната таблица, маркирани в сиво, съдържат отрицателни корелационни коефициенти. Както може да се види, най-силната взаимовръзка между двойка въпроси (65, 67) е негативна (r_{tet} = -0.232), а най-високата позитивна корелация (въпр. 62, 67) е под това равнище (r_{tet} = 0.148).

Наличието на множество ниски, включително и негативни интракорелации без съмнение е белег са слаба вътрешна консистентност на субтестовете като разглеждания. Данните в следващата таблица дават представа за равнищата на интра- и интеркорелациите на въпросите от този субтест. Това са коефициентите на множествена корелация R и на детерминация R^2 на съответния въпрос с всички останали въпроси от субтеста, както и с всички останали въпроси извън субтеста.

Таблица 20. Интра- и интеркорелации между въпросите от вариант 141, субтест 7. *Химия*

	Интракорелации		**Интеркорелации**	
Въпроси	**коефициент на множествена корелация *R***	**коефициент на детерминация R^2**	**коефициент на множествена корелация *R***	**коефициент на детерминация R^2**
61	0.090	0.008	*0.397	*0.158
62	0.127	0.016	*0.446	*0.199
63	0.095	0.009	0.347	0.121
64	0.105	0.011	0.339	0.115
65	0.134	0.018	0.362	0.131
66	0.109	0.012	0.377	0.142
67	0.114	0.013	*0.407	*0.166
68	0.107	0.011	0.374	0.140
69	0.078	0.006	0.353	0.124
70	0.076	0.006	0.342	0.117

Заб. Всички коефициенти, маркирани със звездичка (*), са значими при $\alpha = 0.05$

Сред особеностите на данните в горната таблица следва да отбележим липсата на статистическа значимост на по-голяма част от получените коефициенти на корелация. Няколкото значими статистики обаче са тези, които са използвани като мярка за „външните" корелации на въпросите от този раздел. Изкушаваме се все пак да посочим, че контрастът между равнищата на интра- и интеркорелациите на въпросите от този субтест е дори по-силно изразен, отколкото между тези при въпросите от субтеста по математика, представени в таблица 18.

Впрочем, обратнопропорционалните връзки между въпросите от една и съща предметна област не са рядко изключение. В следващата таблица са представени данни за количеството на негативните коефициенти и техния дял от всички в съответния субтест. Както сочат данните, не само при разгледания по-горе субтест по химия от вариант 141, но и при други субтестове негативните взаимовръзки имат немалък дял. Такива са например субтест 8. *Биология* от вариант 141 (21 коефициента, 46.67%), субтест 6. *Физика* от същия вариант (15, 33.33%) или субтест 7. *Химия* от вариант 134 (14, 31.11%).

Таблица 21. Дял на негативните корелации по тестови варианти и субтестове

Субтест	Вариант 134		Вариант 141		Вариант 171	
	брой	процент	брой	процент	брой	процент
1. Български език	0	0.00	8	17.78	11	24.44
2. Литература	8	17.78	13	28.89	6	13.33
3. История	9	20.00	14	31.11	8	17.78
4. География	6	13.33	11	24.44	14	31.11
5. Математика	5	11.11	5	11.11	0	0.00
6. Физика	11	24.44	15	33.33	9	20.00
7. Химия	14	31.11	21	46.67	14	31.11
8. Биология	13	28.89	21	46.67	8	17.78
9. Разсъждения	1	2.22	6	13.33	9	20.00
10. Семантика	2	4.44	2	4.44	1	2.22

Макар че данните са извлечени от сравнително малък брой тестове, могат да се забележат някои характерни тенденции. С две изключения от представените в таблицата 30 субтеста, всички останали субтестове съдържат въпроси, които корелират, в по-малка или по-голяма степен, негативно с един или повече от останалите въпроси в същия субтест. Тестовите варианти се отличават и по обема на негативните корелации между въпросите – при вариант 141 те са близо два пъти повече от тези в останалите два варианта. Различават се и отделните субтестове, като негативни корелации се наблюдават по-често в субтестовете в областта на природните науки (физика, химия и биология), към които можем да добавим разделите по география и история, отколкото в някои хуманитарни области (български език, разсъждения, семантика). С по-малък обем на негативните корелации се характеризират и субтестовете по математика.

Ще припомним, че субтестовете от първата група се асоциират по-често с ниски, под праговото равнище собствени стойности, с ниска обяснителна сила на еднофакторните им конфигурации и със слаба съдържателната плътност.

Направените анализи представят достатъчно свидетелства, че корелационните връзки между въпросите, принадлежащи на една и съща предметна област, не са особено силни. Нещо повече, тестовите въпроси имат валенции към въпроси от други субтестове, понякога по-силни, отколкото към въпросите от същия субтест. Може, следователно, да се предположи, че факторната структура на цялостния тест ще се различава от тази, която произтича от неговия дизайн. Във всеки случай може да се очаква, че латентната структура не е конституирана на предметен принцип.

2.3.2.3. Определяне на факторните модели

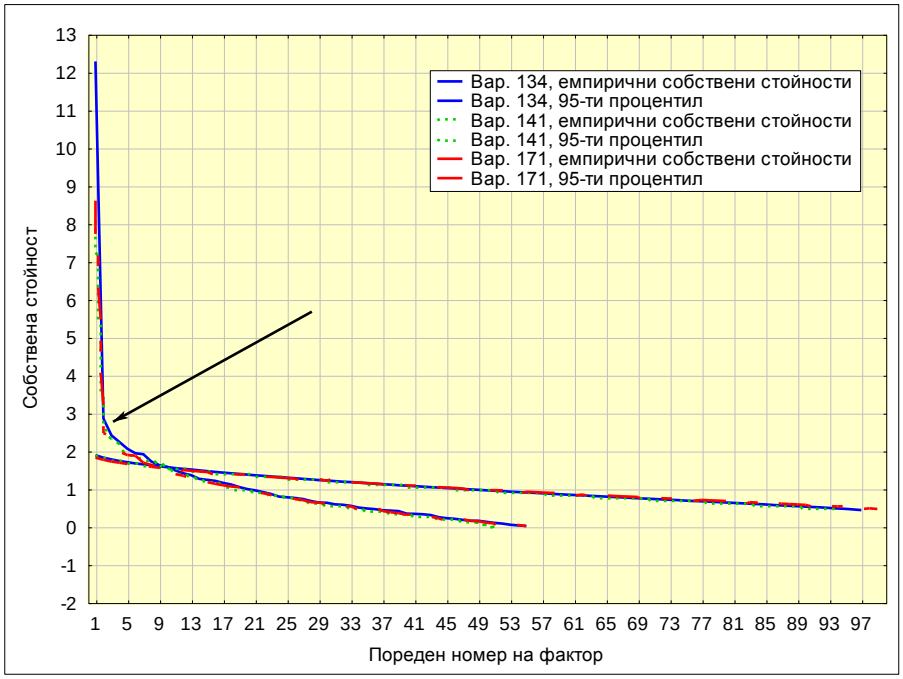
Прилагането на паралелния анализ върху първоначалните факторни конфигурации доведе до съществено редуциране на броя на факторите, които следва да бъдат включени във финалните модели. Така например финалните незавъртени факторни структури при вариант 134 и 141 включват 10 фактора, а при вариант 171 – 8 фак-

тора (виж таблица 17 и приложение 8). При тази процедура от първоначално извлечените латентни конфигурации бяха отстранени 80 – 85% от факторите със собствени стойности, по-ниски от съответните референтни стойности.

Съществото редуциране на първоначалните факторни структури доведе и до съществено намаляване на дела на обяснената споделена дисперсия на въпросите. Кумулативната обяснителна сила на факторите в отделните конфигурации, подобно на ситуацията на субтестово равнище, не може да се разглежда като задоволителна. Така например кумулативният дял от цялата дисперсия на въпросите от вариант 134, който може да бъде обяснен чрез финалния модел, е 31.53% (при 59.56% в първоначалната конфигурация), при вариант 141 този дял е 27.26% (срещу 53.86%), а при вариант 171 - 22.73% (срещу 52.00%). Както се вижда, около 68 – 77% процента от дисперсията на въпросите в тестовите варианти следва да се разглежда като уникална или дължаща се на грешките на измерването.

Не са високи и собствените стойности на първите по ред фактори, както и дяловете от дисперсията в съответните тестови варианти, която може да бъде обяснена чрез тяхното влияние. Така например най-силният, първи по ред фактор в латентната структура на вариант 134 има собствена стойност 12.28 и на него се дължи 12.66% от цялата дисперсия; при вариант 141 съответните стойности са 8.07 и 8.59%, а при вариант 171 - 8.60 и 8.69%.

Фигура 18. Приложение на паралелния анализ върху резултатите от факторния анализ на тестови варианти 134, 141 и 171



Въпреки това на горната графика, на която са представени собствените стойности на факторите от първоначалните конфигурации на трите разглеждани тестови

варианта, както и съответните 95% референтни стойности от паралелния анализ, може да се забележи яркото противопоставяне на първия срещу останалите по ред фактори. При всеки от профилите – твърде сходни, с изключение на равнищата на първите си фактори, лесно може да бъде идентифицирана точката на пречупване.

Тя е при втория по ред фактор и разделя профилите на силно скосена част наляво от нея и на относително по-хоризонтална част, в която нивата на профилите плавно затихват, съдържащи „факторни отломки" по метафоричния израз на Р. Кетел. Ако използваме графичния тест на Р. Кетел като единствен метод за определяне на оптималния брой на факторите, без колебание бихме идентифицирали тези структури като еднофакторни. В подкрепа на това предположение говорят и високите нива на съотношенията между собствените стойности на първия и втория по ред фактор при разглежданите конфигурации, които възлизат на 4.306 при вариант 134, 3.069 при вариант 141 и 3.488 при вариант 171. Тези резултати могат да се разглеждат като аргумент в полза на хипотезата за едномерност на тестовите латентни структури.

2.3.2.4. Анализ на факторните тегла на въпросите от финалните факторни решения. Постигане на прости факторни структури

Преди да подложим тази хипотеза на проверка, нека да разгледаме факторните тегла на въпросите и разпределението им между отделните фактори. Факторните тегла са изчислени след ортогонална ротация на факторите от финалните модели по метода *varimax normalized*, който е един от най-продуктивните и поради това най-често използваните методи за ротация. Целта на „завъртането" на факторната структура в пространството на променливите е да се получат такива фактори, които са по-податливи на интерпретация. Методът *varimax normalized* се основава на идеята за максимизиране на дисперсията на квадратите на нормализираните факторни тегла на всички променливи (редове) по всеки фактор (колона) от факторната матрица. Този подход води до по-ярко изразена „определеност" на манифестираните променливи чрез (а) намаляване на броя на променливите, които имат (приблизително) еднакви тегла по различните фактори и (б) маркиране на всеки фактор с високи факторни тегла на дадени променливи и ниски – на други променливи. Всичко това се вписва в стратегията за постигане на проста факторна структура, представена по-горе в текста (Thurstone, 1935, 1947; Kim & Mueller, 1978; Reynolds & Kamphaus, 2003). Пълните резултати от ортогоналната ротация на факторите от конфигурацията на вариант 134 са представени в приложение 10.

Матриците на факторните тегла на завъртените конфигурации се характеризират с няколко взаимосвързани особености. Първата от тях е, че собствените стойности на завъртените фактори са по-ниски от тези преди ротацията (при вариант 134 тези стойности са 6.14 на най-силния фактор и 1.98 на най-слабия). Общият обем на споделената дисперсия (установена при финалните незавъртени конфигурации) е „преразп-

ределен" така, че всеки от завъртените фактори обяснява изключително малък дял от него. За двата посочени по-горе фактора този дял е съответно 6.33% и 2.04%.

Като цяло въпросите във факторните матрици се характеризират с ниски факторни тегла. Така например въпросите с тегла, по-ниски (по абсолютна стойност) от 0.10, са около 31% във фактор 1 и около 65% във фактор 4. В останалите фактори делът на въпросите с такива факторни тегла се простира между тези две стойности.

Разпределението на айтемите между завъртените фактори може да се разглежда като недвусмислено потвърждение на предположението, че латентните структури на ТОП не са организирани на тематичен (субтестов) принцип. В следващата таблица е представен броят на въпросите, които, въз основа на тяхното максимално факторно тегло, са отнесени към съответния фактор. Лесно се забелязва, че въпросите от всички раздели на теста се разделят на субгрупи, повлияни от различни фактори. По-компактни са субтест 1. *Български език*, чийто въпроси се влияят от 2 фактора (фактор 3 – 8 въпроса и фактор 1 – 2 въпроса), 5. *Математика*, при който повече от половината въпроси са групирани във фактор 1, както и 9. *Разсъждения*, при който 7 въпроса принадлежат на същия фактор.

Таблица 22. Разпределение на въпросите между факторите от вариант 134 (първични ортогонални фактори)

Субтест	Фактор									
	F 1	F 2	F 3	F 4	F 5	F 6	F 7	F 8	F 9	F 10
1. Български език	2	-	8	-	-	-	-	-	-	-
2. Литература	-	-	2	-	-	2	1	-	5	-
3. История	-	-	-	1	-	1	1	-	6	1
4. География	-	2	-	1	-	4	-	1	2	-
5. Математика	6	1	-	-	1	1	1	-	-	-
6. Физика	2	2	1	-	-	-	3	1	-	1
7. Химия	-	2	2	1	1	1	2	1	-	-
8. Биология	2	-	1	1	-	1	-	2	2	1
9. Разсъждения	7	-	1	-	1	-	-	-	1	-
10. Семантика	4	-	-	-	3	1	1	-	-	1
Общ брой на въпросите	23	7	15	4	6	11	9	5	16	4
Собствени стойности	6.14	2.13	3.86	2.11	2.84	3.15	2.64	1.98	3.45	2.67
Дял от цялата дисперсия	6.33	2.20	3.98	2.18	2.93	3.24	2.73	2.04	3.55	2.34

И обратно, дифузен характер проявяват субтестове като 4. *География*, чийто въпроси са разпределени сравнително равномерно между 5 фактора, 6. *Физика* – между 6 фактора, 7. *Химия* – между 7 фактора и 8. *Биология* – също между 7 фактора.

Извлечените фактори също се различават по броя на айтемите, на които оказват влияние. Така например фактор 1 е маркиран с най-високи факторни тегла по 23 айтема, докато фактор 10 – само по 4. Без съмнение, има зависимост между обема на

айтемите, повлияни от съответния фактор, и неговата собствена стойност – коефициентът на корелация между тези две величини по данните от таблица 22, е 0.92 при $p < 0.05$

По-интересно е да се проследи каква е съдържателната „плътност" на факторите, определена чрез обема на въпросите със „значими" факторни тегла. Ако приложим правилото, възприето в анализа на факторните структури на субтестово равнище, за „значими" да се разглеждат факторните тегла, по високи от ±0.45, структурата на факторите изглежда по начин, представен в следващата таблица. От данните в нея се вижда, че дори и при тази невисока прагова стойност, въпросите със „значими" факторните тегла формират малък дял в структурата на отделните фактори, което, без съмнение, би затруднило тяхната интерпретация.

Таблица 23. Въпроси с факторно тегло над ±0.45 за факторите от завъртяната конфигурация на вариант 134

Фактор	Брой въпроси с тегло > ±0.45	% от всички въпроси	Собствена стойност
F1	8	8.25	6.14
F2	3	3.10	2.13
F3	2	2.06	3.86
F4	1	1.03	2.11
F5	2	2.06	2.84
F6	3	3.10	3.15
F7	2	2.06	2.64
F8	1	1.03	1.98
F9	1	1.03	3.45
F10	1	1.03	2.67

Друга изключително важна особеност е тази, че по-голяма част от променливите имат сравнително високи, в някои случаи почти равностойни факторни тегла по няколко (нерядко - до 4–6) фактора. Ето няколко примера:

- въпрос 3 от раздела по литература, с факторно тегло 0.429 (по фактор 1) и 0.376 (по фактор 3);

- въпрос 9 от раздела по история, с факторни тегло 0.326 (по фактор 3), 0.289 (по фактор 5), 0.389 (по фактор 9);

- въпрос 10 от раздела по география, с факторно тегло 0.452 (по фактор 6) и 0.438 (по фактор 9);

- въпрос 3 от раздела по физика, с факторно тегло 0.197 (по фактор 1), 0.289 (по фактор 2), 0.292 (по фактор 3), -0.204 (по фактор 4), -0.225 (по фактор 5), 0.251 (по фактор 7)

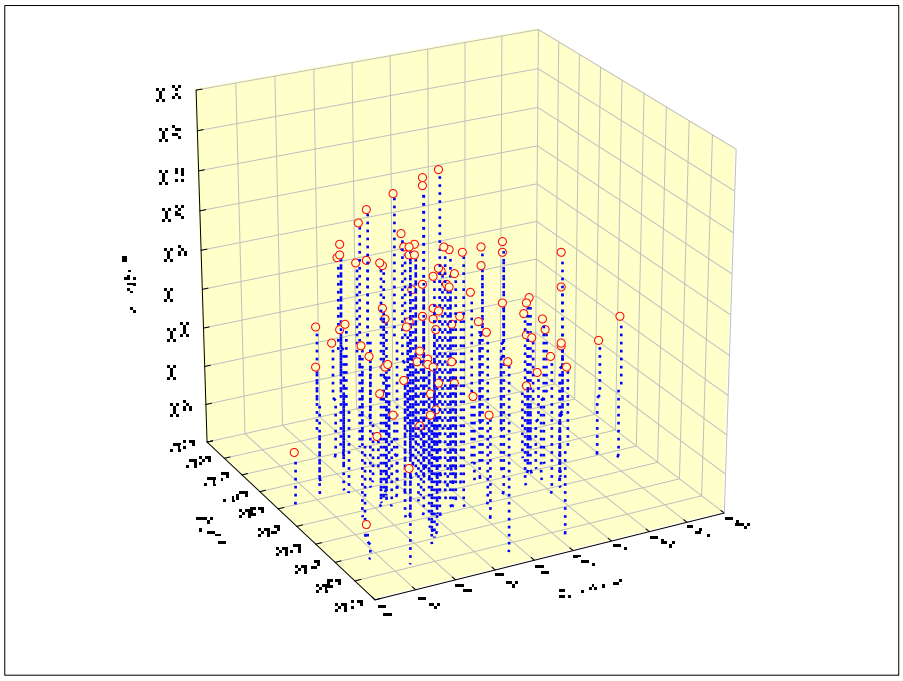
- въпрос 6 от раздела по разсъждения, с факторно тегло 0.247 (по фактор 1),

0.181 (по фактор 3), 0.214 (по фактор 6), 0.265 (по фактор 9).

Трябва да се подчертае, че преобладаващата част от въпросите в ТОП се характеризират с относително високи факторни тегла поне по два от факторите и че обратното е по-скоро изключение. Може да се обобщи, следователно, че отговорите на и. л. на въпросите от ТОП се обуславят от съвместното влияние на няколко фактора и поради това могат да бъдат квалифицирани като многомерни въпроси.

Такива въпроси, които корелират в приблизително еднаква степен с два или повече фактори, обикновено се разглеждат като „мостове" между съответните фактори. Би следвало да очакваме обаче, че след приложения факторен анализ с последваща ортогонална ротация на факторите, латентната конфигурация ще бъде по-отчетлива, с ясно разпределение на въпросите между независимите фактори. Наличието на множество въпроси-мостове обаче е свидетелство, че желаният ефект по-скоро не е постигнат. За това свидетелства и следващата 3D графика, представяща факторните тегла на въпросите по три от факторите с най-високи собствени стойности: фактор 1 (със собствена стойност 6.14), фактор 3 (3.86) и фактор 9 (3.45).

Фигура 19. Тримерна графика на факторните тегла по фактор 30, 31 и 32 на завъртяната конфигурация на вариант 134



Ясно се вижда облакът от точки в централната част на графиката, представящи факторните тегла на отделните въпроси. Той е по-скоро компактен, отколкото разпределен на групи между отделните фактори. Тази конфигурация подсилва впечатлението за хомогенността на въпросите от гледна точка на тяхната факторна принадлежност.

Въз основа на направените анализи на факторните тегла на айтемите може да

се направи предположението, че ортогоналната ротация не е довела до свеждането на незавъртените факторните модели до по-прости (чисти) структури и че тези структури едва ли биха могли да получат смислена интерпретация.

Значението, което Л. Л. Търстоун влага в това понятие, е дали завъртените решения са единствени, уникални и най-добри от гледна точка на възможностите за тяхната интерпретация (Thurstone, 1934; 1935; 1936). Както бе отбелязано по-горе (виж т. 1.2.3.4.), Л. Л. Търстоун поставя набор от пет ограничения (условия) към факторния модел, спазването или по-скоро постигането на които би довело до възможно най-простия, но смислен и теоретично съдържателен модел на връзките между манифестираните и латентните променливи.

Ако условията на Л. Л. Търстоун бъдат приложени върху факторната матрица на вариант 134 от приложение 10, ще се получат следните резултати:

1. Всеки ред от факторната матрица трябва да съдържа поне една нулева стойност – на това условие отговарят само 3 въпроса: 31 (раздел 4. *География*) по фактор 3, 34 (раздел 4. *География*) по фактор 6 и 99 (раздел 10. *Семантика*) по фактор 7.

2. Ако има m общи фактори, във всяка колона от факторната матрица трябва да има поне m нулеви стойности – нито една колона не отговаря на това условие

3. Във всеки две колони от факторната матрица трябва да има няколко променливи, чиито факторни тегла клонят към нула в едната колона, но не и в другата – всички двойки колони отговарят на това условие.

4. Във всеки две колони от факторната матрица, голяма част от променливите трябва да имат клонящи към нула стойности в двете колони, ако има 4 или повече фактора – нито една от двойките колони не отговаря на това условие.

5. Във всеки две колони на факторната матрица трябва да има само малък брой променливи с неклонящи към нула стойности в двете колони – малък брой двойки колони отговарят на това условие

Последователното „налагане” на общите условия на Л. Л. Търстоун върху матрицата на факторните тегла на айтемите разкрива, че данните не се съгласуват с по-голяма част от тези условия. Особено показателна е несъгласуваността с първото (основно) условие, на което отговарят само няколко променливи, както и с второто условие. По-висока е степента на съгласуваност с условия 3 и 5, особено с третото, което се отнася за отделни двойки колони. Като цяло обаче по-голяма част от въпросите се характеризират с близки (ненулеви) тегла по няколко фактора. Примери за такива променливи бяха представени по-горе в текста.

Тези обстоятелства водят до нарушаване не само на изходните пет условия на Л. Л. Търстоун, но и на по-опростеното условие на Р. Торндайк, съгласно което максимална простота на факторната структура е постигната тогава, когато дадена променлива има определено (високо) тегло само по един от факторите и нулеви тегла по

всички останали фактори (Thorndike, 1971).

Резултатите от направените анализи на завъртяната факторна структура, съставена от ортогонални фактори, води към извода, че постигнатото решение не покрива условията за проста факторна структура. По всичко изглежда, че по-подходящо би било да се приложи неортогонална ротация на факторите от избрания модел.

2.3.2.5. Йерархичен факторен анализ на избраното решение

Използваната в предходния анализ ортогонална ротация на факторите е процедура, която води до решения с независими, некорелиращи фактори. Алтернативна на тази процедура е такъв тип завъртане на факторното пространство, което „позволява" на факторите да корелират. Концепцията за неортогоналните (наклонени) фактори е често експлоатирана в изследванията в областта на социалните и хуманитарните науки, тъй като тук е нормално да се очаква, че латентните димeнсии не са изолирани една от друга. Методите с използване на наклонени фактори се прилагат тогава, когато ортогоналните методи не са довели до проста, лесна за интерпретация факторна структура. В тези случаи неортогоналната ротация би могла да доведе до по-точна, по-възпроизводима факторна структура (Harman, 1976; Costello & Osborne, 2005).

През този етап на изследването е използван йерархичният факторен анализ, който представлява развитие на "конвенционалната" методология на въртене на наклонени фактори. Йерархичният факторен анализ е многостъпкова процедура, която започва с идентифициране на групи (клъстъри) от променливи с високи интеркорелации. Факторите се завъртат така, че да представят тези клъстъри по най-добрия начин, без ограничение за ортогоналност. Изчислява се корелационната матрица на наклонените фактори, която се подлага на вторично факторизиране, за да извлече финална структура от ортогонални фактори, които разделят дисперсията на наблюдаваните променливи на две части: дисперсия, която се дължи на влиянието на общите фактори (фактори от втори ред) и уникална дисперсия, която се дължи на клъстърите от променливи (фактори от първи ред).

Чрез процедурите на йерархичния факторен анализ бяха извлечени 10 фактора от първи ред и 5 фактора от втори ред. Корелациите между наклонените фактори (клъстъри от променливи) варират от -0.03 (между фактори 2 и 10) до 0.66 (между фактори 3 и 9), по-голяма част от които са умерено високи. По-силно изразени са корелациите между клъстърите от променливи и факторите от първи ред, представени в приложение 11. Те се изменят между 0.59 (клъстър 3 и фактор 3) и 0.90 (клъстър 9 и фактор 9). Тези данни са свидетелство, че между наклонените фактори се наблюдава определена корелация, както и между клъстърите от променливи и факторите от първи ред, което подхранва убедеността, че приложеният метода на ротация е адекватен.

Ние ще фокусираме вниманието си към общите фактори от втори ред, които генерират общата (споделена) дисперсия. В същото приложение са представени и коре-

лациите между клъстърите от променливи и факторите от втори ред. Вижда се сравнително отчетливото разпределение на клъстърите от променливи между вторичните фактори. Наблюдават се и групови крос-корелации, т. е. един и същи клъстър корелира с повече вторични фактори, какъвто е например клъстър 9, които е обвързан последователно с всички фактори от втори ред с корелации съответно 0.239, 0.225, 0.399, 0.598, 0.153. Можем да очакване, че и при матрицата на факторните тегла на въпросите по факторите от втори ред ще се наблюдават такива крос-корелации. Факторните тегла на въпросите от вариант 134 с извлечените 5 фактора от втори ред са представени в приложение 12. В съгласие с очакванията, голяма част от въпросите имам приблизително еднакви факторни тегла по няколко (поне два) от вторичните фактори. Като пример ще посочим въпрос 30 от субтест 3. *История*, който корелира с първите 4 фактора с тегла съответно 0.23, 0.14, 0.16 и 0.22. Същевременно факторните тегла и при този анализ са сравнително ниски, вариращи от -0.55 (въпрос 22 по фактор 4) до 0.49 (въпрос 97 по фактор 3). Ето какво е групирането на въпросите по фактори:

Таблица 24. Разпределение на въпросите между факторите от втори ред от вариант 134

Субтест	Фактори				
	F 1	F 2	F 3	F 4	F 5
1. Български език	6	-	4	-	-
2. Литература	5	-	5	-	-
3. История	5	-	2	3	-
4. География	1	-	3	5	-
5. Математика	3	1	2	3	-
6. Физика	4	2	1	3	-
7. Химия	-	2	2	6	-
8. Биология	3	1	2	4	-
9. Разсъждения	5	-	3	1	-
10. Семантика	3	1	5	1	-
Общо:	35	7	29	26	0

Забележка: Броят на въпросите в субтестовете по география, математика и разсъждения, включени в корелационната матрица, е 9.

Независимо от сравнително ниските факторни тегла и наличието на крос-корелации, разпределението на въпросите между факторите според най-високото им факторно тегло позволява да се направи смислена интерпретация, поне на някои от тях. Начинът, по който въпросите са групирани в отделните вторични фактори, още веднъж потвърждава тезата,че факторната структура на ТОП не е изградена на тематичен принцип. Въпросите от различните субтестове са разпределени в 2 до 4 фактора от втори ред. С най-голям брой въпроси се характеризират фактори 1, 3 и 4. Последният, пети по ред фактор, не е маркиран с високо факторно тегло по нито един от въп-

росите и поради това ще бъде изключен от анализа. Фактор 3 влияе върху 7 въпроса, които имат ниски факторни тегла, със съдържание, което също няма ясна интерпретация. Ето защо тук ще представим оставащите 3 фактора от втори ред.

По-голяма част от въпросите, които гравитират към фактор 1, изискват способности за прилагане на по-общи правила, норми, закономерности или принципи в определени, конкретни ситуации. По-често тези правила са част от общите компетентности в съответната предметна област и не се съдържат експлицитно в основата на въпроса. От този вид е въпросът по математика, даден като пример по-долу. Формално, това е задача за прилагане на едно аритметично действие, но прилагането на друго правило би оптимизирало нейното решаване. Някои въпроси „подканват" и. л. да приложат определено неназовано правило, какъвто е въпросът по български език, в чиято основа е посочена областта на правописните правила. При някои въпроси правилото е, повече или по-малко, експлицирано чрез използването в основата на въпроса на релевантни термини или символи, каквито са примерните въпроси по физика и биология. И накрая, в някои въпроси, главно от разделите по разсъждения и семантика, правилото е заложено в основата на въпроса (в някои случаи – с подкрепата на алтернативните отговори), а от и. л. се очаква да го изведат и/ или приложат.

Въпросите с относително високи факторни тегла по този фактор, който може да бъде определен като „способност за използване на правила", са предимно от субтестовете по български език, история, математика, биология, разсъждения и семантика. Ето няколко примера на въпроси с високи тегла по този фактор.

• Субтест 1. *Български език* (факторно тегло 0.41)

Кое прилагателно е написано правилно?

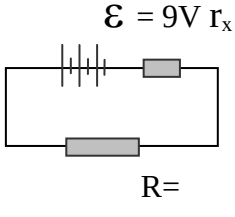
а) искренно признание
б) наследственна болест
в) безсмисленни действия
г) откровенни признания
д) едновременни събития

Правилото, което трябва да бъде приложено, е за правопис на двойно „н".

• Субтест 6. *Физика* (факторно тегло 0.27)

Каква е големината на вътрешното съпротивление на източника, ако I = 2A?

а) 0,5 Ω
б) 0,3 Ω
в) 2 Ω
г) 5 Ω
д) 1,5 Ω



Тук от и. л. се очаква да активизират знанията си за електродвижещо напрежение и вътрешно съпротивление на източник и да приложат към конкретната схема за-

висимостта между няколко физични величини, известна като Закон на Ом за затворена електрична верига, която се представя с конкретна формула.

• Субтест 10. *Семантика* (факторно тегло 0.27)

Слепота се отнася към *цвят*, както *глухота* към...

а) слух
б) музика
в) звук
г) думи
д) гама

В този въпрос от и. л. се очаква да установят подобие на две релации от типа a:b = c:d. За да направят това, най-напред трябва да установят типа на релацията между първите две думи (да изведат правило), за да изградят втората релация.

• Субтест 8. *Биология* (факторно тегло 0.27)

Далтонизмът е наследствена болест, свързана с пола. Дължи се на рецесивен алел в Х-хромозомата. Ако родителите имат нормално зрение, но синът им е далтонист, какъв е генотипът на тримата?

	МАЙКА	*БАЩА*	*СИН*
а)	X^AX^a	X^AY	X^AY
б)	X^AX^a	X^aY	X^aY
в)	X^aX^a	X^AY	X^aY
г)	X^AX^A	X^AY	X^aY
д)	X^AX^a	X^AY	X^aY

За да се отговорят на този въпрос, и. л. следва да знаят още, че далтонизмът се причинява от рецесивен алал на ген, който липсва в Y-хромозомата и че заболяването се предава от майка на син. След това да приложат правилото за унаследяване в конкретната ситуация.

• Субтест 5. *Математика* (факторно тегло 0.27)

На колко е равна сумата на числата от 1 до 10 000?

а) 50 005 000
б) 10 000
в) 50 000 000
г) 10 001
д) 49 995 000

Задачата изглежда изчислителна, но тук трябва да се приложи правилото за изчисляване на сумата на първите *n* члена на аритметичната прогресия, което се изразява с конкретна формула.

По-голяма част от въпросите, които конституират фактор 3, изискват способнос-

ти за обобщаване на информацията, за нейното синтезиране или генерализиране. В някои случаи обобщението е на по-ниско ниво, например да се намери общото (сечението) между два обекта, какъвто е примерният въпрос по семантика. В други случаи е необходимо да се направи обобщаваща оценка на даден феномен (примерният въпрос по география) или да се извлече общото от множество феномени (примерните въпроси по литература и математика). Интересни са въпросите, при които задачата е да се отнесе даден обект към група от обекти със сходни свойства (задача за класифициране), например при изясняване на даден правописен проблем, както в примерния въпрос по български език. Този айтем не е за прилагане на правописно правило – при правописа на начално о/ у няма формални правила. Правописът се определя от семантиката на думата (напр. действието на глагола се довежда до резултат) и от нейната етимология, така че и. л. трябва да отнесе съответната дума към по-широка категория от думи със съответната семантика и етимология, за да установи правописната норма.

Въпросите с относително високи факторни тегла по този фактор, който може да бъде определен като „способност за обобщаване", са предимно от субтестовете по хуманитарните дисциплини.

Ето няколко примера на въпроси с високи тегла по този фактор.

• Субтест 10. *Семантика* (факторно тегло 0.49)

С коя сричка могат да се образуват две отделни думи, ако сричката е край на първата и начало на втората дума от следната двойка:

па… ; …*ка*

а) лат
б) топ
в) нер
г) лет
д) прат

• Субтест 4. *География* (факторно тегло 0.34)

Туризмът е един от най-перспективните отрасли на националната икономика, защото…

а) е източник на валутни постъпления
б) стимулира развитието на транспорта, строителството и търговията
в) осигурява временна и постоянна трудова заетост
г) осигурява бърза възвращаемост на вложените инвестиции
д) по всички изброени причини

• Субтест 10. *Семантика* (факторно тегло 0.30)

Отбрана, *врата* и *топка* обобщете с…

а) хокей
г) бейзбол

б) баскетбол
в) футбол
д) крикет

• Субтест 2. *Литература* (факторно тегло 0.29)

Образът на дявола се среща в творчеството на...

а) Христо Смирненски
б) Гео Милев
в) Николай Лилиев
г) Веселин Ханчев
д) Дора Габе

• Субтест 1. *Български език* (факторно тегло 0.26)

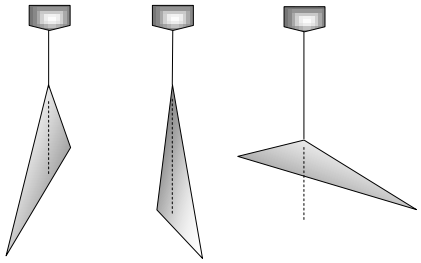
В коя дума има правописна грешка?

а) оцветявам
б) ограмотявам
в) обменям
г) овеселявам
д) обогатявам

• Субтест 5. *Математика* (факторно тегло 0.26)

Ако окачваме последователно върховете на един и същ произволен триъгълник на конец, то мислените линии, продължаващи от конеца при всяко окачване, ще се пресекат в една и съща точка, лежаща на триъгълника. За кое от изброените тела ще се получи същото?

а) куб
б) призма
в) тетраедър
г) пирамида
д) за всяко от изброените тела



Въпросите, формиращи фактор 4, без колебание могат да бъдат класифицирани като въпроси за възпроизвеждане на знания. Тяхното съдържание е свързано с припомняне на термини (чрез техните дефиниции), дефиниране на понятия, посочване на факти, характеристики, обстоятелства и явления, различаване на обекти, както и установяване и прилагане на прости правила. Не е случайно, че най-голям дял по този фактор имат природните дисциплини, както субтестовете по история и география. Можем да определим този фактор като „възпроизвеждане на знания".

• Субтест 7. *Химия* (факторно тегло 0.43)

Известно е, че 1 mol от всяко вещество съдържа 6,02.1023 частици. Как се нарича това число?

а) константа на Планк
б) число на Фарадей
г) универсална газова константа
д) молекулна маса

в) число на Авогадро

• Субтест 8. *Биология* (факторно тегло 0.41)

От изброените кости определете коя не е кост на свободния горен крайник.

а) лакътна кост
б) лъчева кост
в) гръдна кост
г) раменна кост
д) кости на китката

• Субтест 5. *Математика* (факторно тегло 0.40)

Кое е следващото число в редицата: 16, 12, 17, 13, 18, 14, 19, …

а) 15
б) 16
в) 17
г) 21
д) 22

• Субтест 4. *География* (факторно тегло 0.34)

Растениевъдството в Източния Тракийско-Родопски регион е специализирано в областта на...

а) зърнопроизводството
б) овощарството
в) техническите култури
г) лозарството
д) фуражните култури

• Субтест 6. *Физика* (факторно тегло 0.29)

Как се нарича разпадането ${}^{A}_{z}X \rightarrow {}^{A}_{z+1}\mathrm{Y} + {}^{o}_{-1}e + \tilde{\nu}$, като $\tilde{\nu}$ е антинеутрино.

а) α -разпадане
б) γ -разпадане
в) β^{+} -разпад
г) β^{-} -разпад
д) електронно залавяне

2.3.2.6. Потвърдителен факторен анализ

Нека да се върнем към предположението за едномерност на латентните тестови структури, което се подхранва от графичния тест на Кетел, въз основа който може да бъде взето еднозначно решение за едномерност, както и от съотношенията между големините на първите и вторите по ред фактори в незавъртените факторни конфигурации. Резултатите от проверката ще бъдат илюстрирани върху данните от вариант 134, предпочетен поради най-високото съотношение между собствените стойности на първия и втория по ред фактор (4.306).

За проверка на хипотезата за едномерност е използван модулът SEPATH от статистическия пакет STATISTICA. Тестваният модел се проверява посредством специфичен програмен език PATH1, чрез който кодира съответната верижна диаграма.

Приложена е същата методология, както при анализите на субтестово равнище. Като метод за оценка на параметрите на модела е използвана двустъпкова процедура GLS→ML, стартираща с генерализирания метод на най-малките квадрати, последван от метода на максималното правдоподобие. Процедурата е приложена върху същите корелационни матрици с коефициенти на тетрахорична корелация, използвани и при изследователския етап на факторния анализ.

Тук ще обърнем внимание на няколко особености. Три от основните статистики като *MRC*, чрез която се оценява дали итеративната процедура за оценка на параметрите на модела е завършена успешно; *ICSF*, който указва инвариантността на модела по отношение на константен скалиращ фактор и *ICS*, свързан с инвариантността на модела по отношение на промени в скалата, имат нулеви стойности, които са благоприятни по отношение на изградения модел.

Оценките на параметрите на модела, подробни данни за които са представени в приложение 13 показват, че част от факторните тегла на въпросите от анализирания тест по общия фактор могат да бъдат определени като относително високи, но като цяло техните стойности се простират почти равномерно в широкия интервал от 0.014 до 0.665. В допълнение, четири от въпросите имат неголеми отрицателни факторни тегла. Интересно е, че въпросите с най-високи факторни тегла, които биха подпомогнали интерпретацията на фактора, принадлежат към няколко различни предметни области, сред които отсъстват природните науки (физика, химия и биология), но също така и литература и география. Сред първите няколко въпроса с най-високи факторни тегла са тези с номера 41 (математика, факторно тегло 0.665), 3 (български език, 0.638), 29 (история, 0.603), 97 (семантика, 0.596), 82 (разсъждения, 0.563), 84 (разсъждения, 0.559), 6 (български език, 0.531), 45 (математика, 0.528), 90 (разсъждения, 0.508), 26 (история, 0.496) и т. н. Стандартните грешки на оценките са много ниски, нивата на статистическа значимост на проверяващата T-статистика, с няколко изключения, позволяват последователното отхвърляне на хипотезите за нулев стойности на параметрите, което означава оставянето им в модела. Въпросите със статистически незначими коефициенти са 8 и сред тях най-вече са представителите на онези предметни области, които като цяло имат по-ниски факторни тегла – математика (1 въпрос), физика (3 въпроса), химия (3 въпроса) и биология (1 въпрос). Изваждането на тези въпроси от анализа води до слабо подобряване на оценките на неговите параметри, но не толкова, че да промени цялостната картина, за това ще продължим анализа въз основа на първоначалните оценки.

И тук, подобно на резултатите от анализите на субтестово равнище, правят впечатление високите факторни тегла на въпросите по уникалните фактори, които в преобладаващата част имат стойности 0.80 – 1.00, системно надвишавайки тези, оценени за общия фактор. Това би могло да означава, че по-голяма част от цялата дисперсия се дължи на въздействието на уникалните фактори, а не на общия фактор.

Интересни са резултатите от приложението на специфичните критерии и статистики за оценка на годността на проверявания модел. Тук ще отбележим изключително високата стойност на функцията на несъответствията (37.591), която е първото важно свидетелство за неговата неадекватност, както и високата стойност RMS стандартизирания остатък (0.087), която надхвърля критичната стойност от 0.05. Стойността на един от основните тестове за проверка на годността на модела, какъвто е критерият хи-квадрат (26764.557), както и неговата статистическа значимост (p = 0.000), водят към отхвърляне на нулевата хипотеза за пълна адекватност на тествания модел.

Таблица 25. Обобщени резултати от потвърдителния факторен анализ върху данни от вариант 134

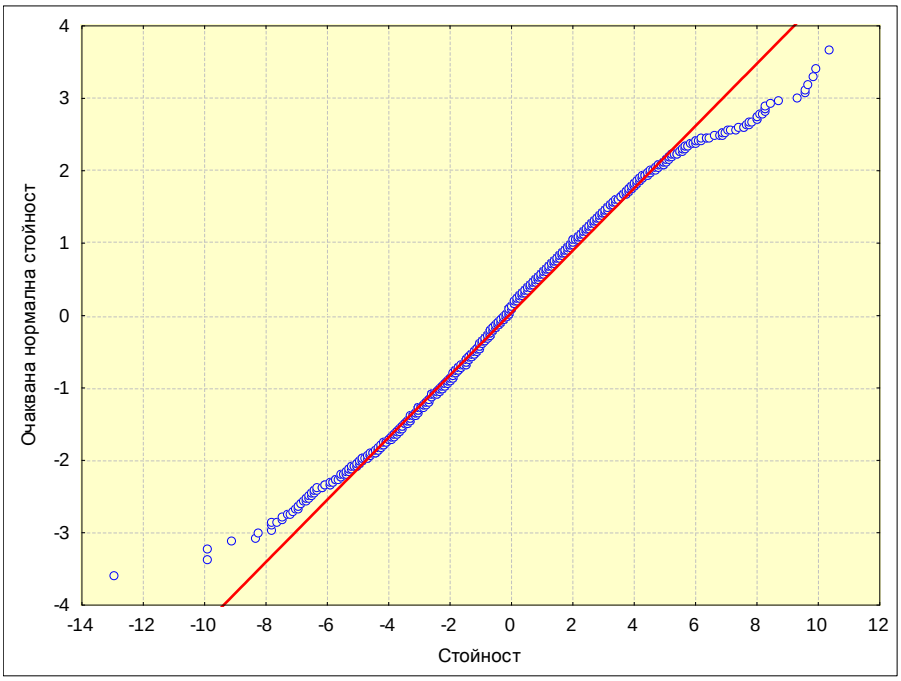
		Статистики			**Приемливи стойности**
Изходни данни		тетрахорични корелации			
Метод за оценка		GLS-ML			
Стойност на функцията на несъответствията		37.591			
RMS стандартизиран остатък		0.087			
χ^2		26764.557			
df		4559			
равнище на значимост		0.000			
едноизвадкови индекси	GFI на Йореског	0.496			>0.95
	AGFI на Йореског	0.474			>0.95
	Индекс на Бентлер-Боне	0.222			≅ 1.00
критерии, базирани на популационния нецентрален параметър		90% долна граница на дов. интервал	Точкова оценка	90% горна граница на дов. интервал	
	Популационен нецентрален параметър	42.124	42.956	43.799	
	RMSEA индекс на Стайгер-Линд	0.096	0.097	0.098	<0.05
	Нецентрален индекс на МакДоналд	0.000	0.000	0.000	>0.95
	Популационен индекс гама	0.525	0.530	0.535	>0.95
	Изравнен популационен индекс гама	0.505	0.510	0.515	>0.95

Далеч извън благоприятните за модела интервали са стойностите на едноизвадковите критерии на К. Йореског и Д. Сьорбом *GFI* и *AGFI*. Особено голямо е несъответствието, което се наблюдава при теста на Бентлер и Боне, чиято стойност клони

към 1.00 при пълна адекватност на модела. Като цяло, едноизвадкови индекси свидетелстват за неадекватността на тествания едномерен модел. В съгласие с тези оценки са и резултатите от тестовете, базирани на популационния нецентрален параметър. Нито точковите, нито интервалните им оценки попадат в областите, означаващи приемлива степен на адекватност на модела (критичните стойности на индексите, показващи висока степен на съгласуваност, са по Steiger, 2009).

Друг аргумент срещу предположението за едномерност предоставя нормална вероятностна графика на нормализираните остатъчни стойности.

Фигура 20. Нормална вероятностна графика на остатъчните стойности при вариант 134



Графиката има сравнително отчетлива S-образна форма, двата края на която видимо се отклоняват от правата линия, като левият й край е над линията, а десния – под нея. Тази характерна форма, според Тюки и неговите колеги, следва да се интерпретира като белег за отклонение на разпределението на остатъчни стойности от нормалното, по-конкретно за наличието на по-„леки” опашки на наблюдаваното разпределение (Hoaglin, Mosteller & Tukey, 1991, стр. 187). Отклонения, макар и по-слаби, се забелязват и във вътрешната зона на графиката, свидетелстващи за отклонения на наблюдаваните честоти непосредствено отляво и отдясно от центъра на разпределението.

2.4. Дискусия

Основната цел на изследването е да се проучи размерността на латентните структури, които лежат в основата на Теста за общообразователна подготовка и обус-

лавят отговорите на и. л. на въпросите, и по-конкретно да се направи проверка валидността на допускането за едномерност на тези структури. Проучването бе направена на две равнища – на равнище субтест и на цялостен тест. Тъй като основният метод за валидизация на това допускане е факторният анализ, преди да обобщим основните резултати от направените изследвания, ще направим разграничение между два основни подхода за тяхното интерпретиране.

Редица автори подчертават разликата между абсолютното и евристичното използване на факторния анализ. Абсолютният подход предполага, че резултатите от едно изследване дават пълно, изчерпателно описание на взаимовръзките между разглежданите променливи. Той е израз на увереността, че разработеният модел е единствено възможен и безалтернативен, че прави най-доброто описание на съответния сегмент от действителността. Обратно, евристичният подход предполага разглеждането на получения модел не като абсолютна истина, а като удобна и приложима форма. Евристичният модел няма претенциите за пълнота и изчерпателност, но може да предложи най-подходящия начин за обобщаване на особеностите на данните (Kim & Mueller, 1978; Darlington, 1997).

В настоящото изследване на анализ са подложени множество съвкупности от тестови данни (на субтестово и тестово равнище). Най-важната обща особеност на данните е тяхната ниска степен на структурираност, клоняща към аморфност; взаимовръзките между айтемите са в недостатъчна степен ясно, силно и последователно изразени, което води до слабо и недостатъчно консистентно проявление на моделите на тяхната организация. Поради това резултатите от направените анализи, макар и да демонстрират известна устойчивост и повторяемост, се характеризират и с немалка вариативност, която в повечето случи представлява ненадеждна основа за извеждане на по-категорични изводи. На тази основа по-подходящият подход за интерпретация на факторните структури би бил евристичният.

Ще припомним няколко особености на проведените факторни анализи, чиято цел е осигуряване на вътрешна (статистическа) валидност на резултатите:

(1) Приложен е методът на главните оси, при който, за разлика от метода на главните компоненти, се отчита общата (споделена) дисперсия на наблюдаваните променливи, което позволява извличането на фактори, които имат съдържателна (психологическа) интерпретация.

(2) За основа са използвани матрици на тетрахоричните коефициенти на корелация, подходящи за бинарни данни.

(3) Търсена е максимална обяснителна сила на изградените факторни модели, съдържащи общи фактори.

(4) Като основен метод за определяне на оптималните факторни структури е използван паралелният анализ на Хорн, който е подходящ за приложение при анализа на главните оси и който позволява идентифициране на фактори, които работят над

случайно равнище.

Тъй като резултатите от анализите на субтестово и тестово равнище се характеризират с редица типологични сходства, тук ще бъде представено обобщение на тези типологични особености, както и съответните специфики

В изследването са представени резултатите от факторния анализ на три тестови варианта (134, 141 и 171), избрани поради това, че са представителни за различни равнища на тетрахоричните коефициенти в съответните корелационни матрици. Като обща стратегия за извличане на последователните фактори на първоначалните (незавъртени) факторни конфигурации е приложен методът за максимизиране на дисперсията във факторите и минимизиране на дисперсията в областите около тях, при хипотеза за 10-факторна структура на субтестово равнище и 100-факторна на равнище тест, при минимална нулева собствена стойност на факторите. Броят на първоначално извлечените фактори е изненадващо голям - при различните субтестове варира между 4 и 7, като по-често срещаните конфигурации включват 5 или 6 фактора (при над 86% от анализираните субтестове). На тестово равнище тяхното количество е още по-поразително – при различните тестови варианти броят на извлечените фактори е между 51 и 55.

Макар че този етап на факторния анализ е все още предварителен, правят впечатление няколко особености. На първо място следва да отбележим големият брой на извлечените фактори в незавъртените решения, който и в двете равнища на анализа възлиза на над 50% от броя на въпросите. По-голяма част от факторите обаче се характеризират с ниски собствени стойности (под 1.00), което е свидетелство за тяхната слаба обяснителна сила. На субтестово равнище обикновено това са факторите след първия по ред, като техният дял обхваща между 75% и 100% от всички фактори в съответната конфигурация. На равнище тест делът на тези фактори е малко по-малък, но също така значителен, около 65 – 66% от извлечените фактори. Изразена като дял от цялата дисперсия, която може да бъде обяснена чрез съответния фактор, тяхна слаба обяснителна сила става още по-очевидна. На субтестово равнище този дял не надхвърля 8%, а на тестово – около и под 1.00% от нея.

Първите по ред фактори, извлечени от отделните субтестови данни, макар и най-ярко изразени, също не се отличават с високи собствени стойности. Те варират между 0.56 и 3.09, генерирайки съответно между 5.56% и 30.86% от цялата дисперсия. На тестово равнище тези гранични стойности са съответно 8.07 и 12.28, като частта от цялата дисперсия на въпросите, която може да бъде обяснена чрез съответните фактори, е между 8.59% и 12.66%.

Известно е, че с увеличаване на броя на факторите се увеличава и обяснителната сила на модела, по-специално се намаляват остатъчните корелации, като с това се намалява и делът на уникалната дисперсия. Независимо от големия брой на факторите в незавъртените решения, тяхната кумулативна обяснителна сила също не може

да се определи като висока. Съвкупният дял от цялата дисперсия в отделните субтестове, обяснена с първоначалните факторни конфигурации, е между 16.28% и 45.73%, като в повече от половината от тях този дял не надхвърля 25%. Тя е по-висока на равнище тест, като варира между 52.00% и 59.56%. Това, разбира се, е споделената дисперсия, което означава, че на субтестово равнище около и над 55% от дисперсията в отговорите на отделните въпроси, а на тестово – около и над 40%, остава необяснена, дължаща се на фактори, уникални за отделните въпроси или на случайни флуктуации. Тези наблюдения са безспорен показател за фрагментарния характер на латентните пространства.

Поради тези особености в данните, прилагането на алтернативни методи за определяне на „оптималния" брой на факторите в съответните конфигурации доведе до несъгласувани, дори противоречиви резултати. Но като цяло броят на факторите във финалните модели беше намален съществено в сравнение с първоначалните им варианти.

На субтестово равнище паралелният анализ на Хорн, предпочетен като водещ метод за изработване на факторните модели, доведе до изненадващи резултати. При част от субтестовите конфигурации този метод дава основание за приемане на еднофакторни модели, с повече или по-малко доминиращ първи фактор. Най-често този модел съответства на данните от субтестовете по български език, математика, разсъждения и семантика. При останалите субтестове, особено тези по природните науки (физика, химия и биология), дори първият по ред фактор обикновено не успява да премине отвъд границата на случайните собствени стойности. Съгласно правилата на възприетия метод, тези конфигурации следва да бъдат характеризирани като конфигурации с „нулева" факторна структура.

Анализът на графичния тест на Кетел за отделните субтестови конфигурации, както и на съотношението между първите два фактора по критериите на Ф. Лорд, водят към същото заключение – по-голяма част от субтестовите латентни пространства могат да бъдат разглеждани като едномерни. Резултатите от потвърдителния факторен анализ, направен върху данни от субтестове, за които, въз основа на предходните методи, може с увереност да се предположи едномерност, водят към противоположното решение. Дори и при тези субтестове, с относително силни, добре изразени първи фактори, хипотезата за едномерност следва да бъде отхвърлена. Дали тези резултати следва да се разглеждат като свидетелство за многомерност на субтестовите структури? Отговорът вероятно е по-скоро положителен, поне за някои конкретни субтестове, които, освен със сравнително по-силен първи фактор, се характеризират и с по-добре изразен втори и дори трети по ред фактор, както сочат данните от графичния тест на Кетел. За други конкретни тестове, с незначителна обяснителна сила дори и на първия по ред фактор, отговорът е по-скоро отрицателен. Аргументация за това може да бъде намерена в общата обяснителна слабост на едномерните модели. В това отношение

данните от изследователския и потвърдителния факторен анализ се съгласуват – на първите по ред фактори се дължи между 5.56% и 30.86% от вариацията в тестовите въпроси, като най-често срещания обем е между 10% и 15%. В тази светлина, ако приемем хипотезата за едномерност на субтестовите латентни структури, следва да се откажем от обяснение на приблизително от 70% - 95% от вариацията, която, във философията на факторния анализ, се дължи на уникалните фактори или на случайни флуктуации.

Следва да добавим, че субтестовите факторни конфигурации се отличават и с различна степен на стабилност/ възпроизводимост. Еднофакторният модел е устойчив при субтестове 1. *Български език*, 5. *Математика*, 9. *Разсъждения* и 10. *Семантика*, макар че обяснителната сила на фактора варира. Този модел е по-слабо възпроизводим при субтестове 2. *Литература*, 3. *История* и 4. *География*, които в около 1/3 от случаите демонстрират „нулева" факторна структура. Факторните структури на субтестове 6. *Физика*, 7. *Химия* и 8. *Биология* са най-малко устойчиви, като в над 2/3 от случите имат „нулева" факторна структура. Може да се каже,че устойчивостта, повторяемостта на факторните модели е по-скоро изключение, отколкото правило. Всички тези наблюдения говорят за структурната слабост, неяснота и нестабилност на субтестовите латентни структури.

Търсенето на инвариантни факторни решения по отношение на определена предметна област е задача, върху която акцентират редица специалисти в областта на изследователския факторен анализ (Yates, 1987). Предполага се, че при сравнително малък брой на въпросите в теста и сравнително голяма извадка от лица, чувствителността на оценените параметри към извадковите флуктуации ще бъде по-малка в сравнение с чувствителността на факторното решение към промени структурата на теста. Може да се каже, че тази особеност е намерила проява в анализираните данни и че намерените факторни решения не са инвариантни на субтестово равнище, а по-скоро динамични.

Въпреки преобладаващо ниските факторни тегла на въпросите, които отслабват съдържателната плътност на извлечените първи фактори и затрудняват, а в някои случаи правят невъзможна тяхната интерпретация, бихме могли да отбележим,че при повечето субтестове първият субтестов фактор може да бъде идентифициран като един от трите, определени на тестово равнище. Обяснението за липсата на стабилност може да бъде потърсено в промените/ разликите в състава на отделните едноименни субтестове, които при всеки тестов вариант съдържат различни тестови въпроси, с различно съдържание. По-силната или по-слаба проява на даден фактор е в зависимост от броя и валидността на въпросите, насочени към измерване на съответната латентна способност. Следователно, факторната структура на субтестово равнище, както и нейната обяснителна сила, са контекстово обусловени, зависими от промените в състава на съответния субтест.

На тестово равнище въз основа на процедурата на паралелния анализ бяха определени финалния конфигурации, включващи 8 – 10 от първоначалните над 50 фактора. Графичните тестове на Кетел обаче водят еднозначно към еднофакторни решения, както и прилагането на критериите на Ф. Лорд. Но резултатите от потвърдителния факторен анализ, подобно на тези на субтестово равнище, водят към отхвърляне на нулевата хипотеза за пълна адекватност на тествания едномерен модел. Тук без колебание можем да приемем допускането за многомерност на латентните структури за валидно, тъй като се основава на резултатите от два формализирани метода. Моделите с 8 – 10 факторна структура, след отстраняването на голям брой, макар и несъществени фактори, губят част от обяснителната сила на първоначалните конфигурации, която се свежда до 22.73 – 31.52% от общата дисперсия на променливите

Тези особености на факторните решения поставят въпроса дали ортогоналното завъртане на факторите на тестово равнище е довело до постигане на прости (чисти) факторни структури в смисъла на Л. Л. Търстоун. Последователното прилагане на условията, определени от него, води към негативно заключение. С други думи, завъртените решения не могат да се разглеждат като единствени, уникални и най-добри от гледна точка на възможностите за тяхната интерпретация.

Чрез неортогонална ротация, проведена по метода на йерархичния факторен анализ, бяха идентифицирани 5 общи фактора втори ред, три от които могат да получат смислена интерпретация. Въз основа на айтемите, които ги конституират, тези фактори бяха определени като „способност за използване на правила”, „способност за обобщаване” и „способност за възпроизвеждане на знания”.

На какво се дължи фрагментарността и обща обяснителна слабост на латентните пространства на общите фактори, която се наблюдава не само при финалните, но и при първоначалните конфигурации? Като конструкции факторните модели се градят на мрежата от корелации между въпросите. Затова е важно да се отбележи, че очакваните високи равнища на вътрешните корелации между въпросите в рамките на отделните субтестове и ниски равнища на корелацията им с въпросите от другите субтестове бяха по-скоро опровергани.

Субтестовете се характеризират със сравнителни ниска степен на съгласуваност между отделните въпроси, нерядко се наблюдават и негативни корелации между отделни въпроси. С други думи, субтестовете се отличават със слаба вътрешна консистентност. Тази особеност се отразява пряко и върху равнището на факторните тегла на въпросите. Като цяло те също не са високи, част от тях са близки до нула. Това, разбира се, означава, че съответните айтеми не са свързани с останалите айтеми от субтестовата скала и/ или са свързани с други фактори, различни от основния.

Особено интересен феномен е наличието на въпроси с отрицателни факторни тегла, които същевременно се характеризират и с ниски или отрицателни дискриминативни индекси. Доколкото интерпретацията на (негативните) факторни тегла и дискри-

минативна сила на айтемите са типологично сходни, тези наблюдения водят не само към хипотезата за възможно типологично сходство между дискриминативния индекс и факторните тегла на въпросите, но и към ново разбиране на психометричното значение на дискриминативната сила като основна характеристика на айтемите. Може да се приеме, че тя отразява степента на съгласуваност на между айтема и съответната латентна променлива, т. е. има отношение към конструктната валидност на равнище тестов въпрос.

Ако интерпретацията на факторите в съответните модели на субтестово ниво премине през процедурата за отстраняване на айтемите с ниски, „незначими" факторни тегла под приетата долна граница, голяма част от тях ще загубят допълнително немалка част от своята обяснителна сила и съдържателна „плътност". Над 1/3 от анализираните тестове следва да бъдат интерпретирани с по-малко от 3 айтема, а при някои от тях интерпретацията би била невъзможна поради липса на айтеми със значими факторни тегла.

Тези особености намират отражение в появата на множество фактори в първоначалните конфигурации, всеки от които оказва влияние върху отделни въпроси или малки групи от въпроси, включително и първият, най-силен фактор, както и в слабостта на еднофакторните модели

Степента на съгласуваност между въпросите на равнище тест не се отличава съществено от тази на субтестово равнище. Най-характерната особеност, на която следва да се обърне внимание, е наличието на относително високи корелации между отделни въпроси или групи от въпроси, принадлежащи на различни субтестове. Тази особеност бе експлицирана чрез извеждането на малък брой общи фактора от първи, а след това и от втори ред, които обединяват въпроси от различни тематични области. Следователно може да се направи изводът, че латентното пространство на равнище тест се различава коренно от формалната му субтестова (предметна) структура.

Друга особеност са относително ниските стойности на факторните тегла спрямо възприетите в литературата препоръчителни долни граници. Сравнително малка част от тестовите въпроси удовлетворяват дори по-либералните минимални прагове. Всичко това засилва впечатлението за слабата обяснителна сила на факторите и поставя под въпрос възможността за тяхното интерпретиране.

Като цяло по-голяма част от въпросите имат относително високо факторно тегло по един от факторите и относително ниско - по останалите, което в повечето случаи позволява идентифицирането на тяхната факторна принадлежност. Същевременно немалка част от въпросите корелират в приблизително еднаква степен с два или повече фактори и служат за „мостове" между тях. Тази особеност се наблюдава при всички видове факторни решения – както при първоначалните незавъртени конфигурации на субтестово и тестово равнище, така и при ортогоналните фактори от първи и втори ред на тестово равнище. Можем да отбележим, следователно, че голяма част от тестовите

въпроси са многомерни и че отговорите на индивидите на всеки такъв въпрос се обуславят от съвместното влияние на различни фактори, образуващи многомерни латентни пространства. От гледна точка на приетото трифакторно латентно пространство на цялостния тест, може да се приеме, че отговорите на голяма част от тестовите въпроси се обуславят от съвместното влияние на факторите от втори ред.

Беше показано, че на субтестово равнище приложените методи за редуциране на броя на факторите в първоначалните конфигурации не водят до еднозначни и съгласувани решения. Може да се приеме, че при *част* от анализираните конкретни субтестове съответната латентна структура се характеризира с наличието на един доминиращ фактор, т. е. тези субтестове могат да бъдат разглеждани като основно едномерни и към тях могат да бъдат приложени определените базови модели. В полза на това решение са резултатите от три (паралелен анализ, графичен тест на Кетел и критерии на Ф. Лорд) от приложените четири алтернативни метода (с изключение на потвърдителния факторен анализ). Поради наблюдаваната неустойчивост на факторните конфигурации, при останалата *част* от анализираните субтестове факторната структура е по-скоро аморфна, без наличието на доминиращ фактор, т. е. с „нулева" факторна структура. Прилагането на базовите, а и на който и да е друг модел, би било безпредметно, доколкото субтестовите резултати не биха могли да се обвържат с конкретна латентна променлива. Тази особеност на част от субтестовете е свидетелство за проблеми, свързани с начина, по който са съставени айтемите, както и с конструирането на съответния субтест. В общия случай, всяка субтестова латентна структура може да бъде описана с един от тези два модела. Може да се очаква, че вероятността съответното латентно пространство да бъде едномерно, е по-висока при субтестовете 1. *Български език*, 5. *Математика*, 9. *Разсъждения* и 10. *Семантика*. По-ниска е тази вероятност при субтестове 2. *Литература* и 3. *История*, а най-малко вероятно да се идентифицира едномерност пространство е при субтестове 4. *География*, 6. *Физика*, 7. *Химия* и 8. *Биология*.

На тестово равнище, след неуспешния опит за постигане на проста факторна структура с първични ортогонални фактори, латентната структура се очертава като включваща три общи способности от втори ред – за използване на правила, за обобщаване и за възпроизвеждане на знания. Тъй като тези латентни способности бяха идентифицирани при всички анализирани тестове, може да се предположи с увереност, че латентната структура, която лежи в основата на Теста по общообразователна подготовка, е трифакторна. Следователно базовите модели, които предполагат едномерност, са неприложими на равнище тест.

Втора част: Съпоставително изследване на проявите на очакваните свойства на теоретичните модели в тестовите данни

1. Обща постановка на изследването

1.1. Цели

В предходните глави бяха посочени няколко важни особености на двете конкурентни теории, свързани със статистиките на тестовите въпроси. CTT и IRT си съперничат и в този аспект, приписвайки един и същи набор от характеристики на тестовите въпроси – трудност, дискриминативна сила и налучкване на правилния отговор. Разбира се, двете теории не са „изоморфни" по отношение на начина, по който описват тестовите въпроси. Те се различават по отношение на броя на дескрипторите (списъкът с характеристики на въпросите при новата теория е по-дълъг и включва още два параметъра), има разлики между съдържанията на съответните понятия (при новата теория те са по-сложни и формализирани), различават се и по математическите подходи за тяхното оценяване. Като цяло обаче тези три характеристики имат сходни интерпретации, съответно функции в процеса на конструиране на психометричните инструменти за оценяване.

Тук трябва да уточним, че в тази част на изследването ще се занимаем преди всичко с това какво е очакваното „поведение" на посочените тестови статистики, когато въпросите са поставени в различни условия, и дали наблюдаваното поведение на тези статистики се отклонява от предвиденото. Тези очаквания се основават на дефинициите на съответните характеристики на въпросите в рамките на съответния теоретичен модел, както и на начините за изчисляване на техните стойности. Ще бъдат разгледани и въпросите, свързани със съвместното вариране на едноименните и на разноименните статистики на въпросите, поставени в едно и също условие.

Класическата тестова теория предлага такива методи за оценка на индексите на трудност и дискриминативна сила на въпросите, които биха могли да доведат до зависимост на получените стойности от извадката от изпитани, въз основа на която са изчислени. Такъв тип зависимост е характерна и за резултатите от теста (наблюдавания тестов бал), чийто стойности са подвластни на инструмента за измерване, чрез който са получени. С други думи, може да се очаква, че индексите на въпросите в рамките на CTT са нестабилни и вариативни при многократно оценяване върху различни извадки от и. л. Обратно, методите за оценка на параметрите на въпросите в рамките на Теорията за отговор на тестов въпрос предполагат независимостта на параметрите от конкретната извадка от изпитани, въз основа на която са оценени. Може да се очаква,

следователно, че оценките на параметрите са стабилни, инвариантни при многократно оценяване върху различни извадки от и. л.. Оценката на личностовия параметър Θ също е независима от инструмента, чрез които е получена.

Например ако с даден тест бъде изпитана група от лица, които имат високи равнища на дадена способност, може да се очаква, че въпросите за тази група ще бъдат лесни, т. е. трудността на въпросите в този тест, определена по СТТ, като цяло ще бъде сравнително ниска. Това предположение се основава на дефиницията, съответно на начина на изчисляване на стойностите на този индекс – като отношение между броя на лицата, отговорили правилно на даден въпрос, и общия брой на лицата, отговаряли на този въпрос. Ако групата от изпитани лица е силна, може да се предположи с основание, че делът на правилните отговори ще бъде висок, т. е. трудността на въпросите ще бъде ниска. Ако със същия тест бъдат изпитани лица с ниски равнища на способности, трудността на въпросите като цяло ще бъде висока.

Поради начина на нейното изчисляване в рамките на СТТ, дискриминативната сила на въпросите също е зависима от извадката от и. л. Конкретните й стойности се определят от дяловете на лицата в силната и в слабата група, отговорили правилно на съответния въпрос. Тези дялове биха могли да се варират, при това значително, при различни извадки от изпитани с различни равнища на способности, особено при нехомогенни извадки.

Такива флуктуации на тестовите статистики биха били невъзможни, ако са определени по методите на IRT, т.е. статистиките на въпросите са инвариантни по отношение на извадката от и. л., по-точно от разпределението на способностите Θ в нея (Hambleton, Swaminathan & Rogers, 1991; Fan, 1998; Baker, 2001; Rasch, 2001). Тази особеност произтича от вероятностния подход за изчисляване на стойностите на параметрите и прилаганата процедура на максималното правдоподобие. Авторът на популярния еднопараметричен „Раш” модел Г. Раш отбелязва, че разпределенията на параметрите в този модел са независими и поради това оценката на трудността на въпросите няма да бъде повлияна от това какви стойности на Θ имат индивидите (Rasch, 2001).

Да разгледаме следния пример. Ако характеристичната крива на даден въпрос бъде построена въз основа на извадка, характеризираща се с ниски когнитивни способности, изграждането на тази крива ще бъде направено само за онази част от нея, която обхваща дела на правилните отговори на лицата от тази група, т. е. частта от характеристичната крива над левия край на скалата Θ. Ако от същата популация бъде формирана друга извадка с високи когнитивни способности, изграждането на характеристичната крива ще бъде направено само за онази част от нея, която обхваща дела на правилните отговори на лицата от тази група, т. е. частта от характеристичната крива над десния край на скалата Θ. И в двата случая оценките на съответните параметри на кривата ще бъдат равни. Това дава основание да се мисли, че „стойностите пара-

метрите са характеристики на въпроса, а не на групата, отговорила на въпроса" (Baker, 2001, стр. 55). Авторите, включително и цитираният, все пак оставят вратата отворена, отбелязвайки, че въпреки че действителните стойности на параметрите са инвариантни, техните оценки могат да варират в различните извадки, но в много тесни граници, оставайки почти равни.

Тези особености на статистиките на въпросите се разглеждат като значителен недостатък на Класическата и важно предимство на Теорията за отговор на тестов въпрос. Те обикновено служат и за основания при избора на теоретичен модел за решаване на различни научни или приложни задачи.

Ето защо основната цел на това емпирично изследване е да се потърсят свидетелства за устойчивостта на „поведението" на статистиките на въпросите от Теста по общообразователна подготовка. С други думи, да се установи дали и в каква степен статистиките на въпросите са стабилни, устойчиви, инвариантни по отношение на извадките от и. л., въз основа на които са получени, или варират при оценка в условията на различни извадки. Поради очакваните различия в поведението на статистиките, обусловено от принадлежността им към една или друга теоретична рамка, тази основна цел ще бъде разложена на три подцели:

1. Да се изследва инвариантността (стабилността) на статистиките на въпросите, определени в рамките на CTT и IRT, в различни условия, т. е. при различни извадки от и. л.

2. Да се изследват (а) взаимовръзките между индексите на въпросите, определени в рамките на CTT, и (б) взаимовръзките между параметрите, определена в рамките на IRT, в едно и също условие, т.е. при една и съща извадка от и. л.

3. Да се изследва съгласуваността между индексите, определени в рамките на CTT, и съответните им параметри, определена в рамките на IRT, в едно и също условие, т.е. при една и съща извадка от и. л.

4. Като неосновна, допълнителна цел на изследването може да се очертае анализа на типа на скалата, образувана от суровите стойности на индекса на трудност(p) на въпросите, изчислени според CTT.

В анализите няма да бъде включен индексът на налучкване на правилния отговор по CTT. Обикновено той се прилага към въпроси с множествен избор и се изчислява като реципрочна стойност на броя на дистракторите. В теста по общообразователна подготовка всички въпроси са от този вид, с еднакъв брой на дистракторите, поради което при всички въпроси този индекс е с константна стойност, равна на 0.20.

1.2. Основни допускания

В съгласие с определените цели на изследването и очертаните по-горе очаквания за степента, в която различните статистиките на въпросите, определени съгласно двата теоретични модела, са податливи на изменения в зависимост от извадката, ще

направим следните основни допускания:

(1) По отношение на стабилността (инвариантността) на тестовите статистики

(а) Допускаме, че стойностите на индексите на трудност (p) и на дискриминативна сила (D, r_{bis}), определени в рамките на СТТ, са зависими от извадките, въз основа на които са получени, и поради това тези индекси се характеризират с нестабилност и вариативност. Нестабилността на индексите беше обоснована по-горе в текста.

(б) Допускаме, че стойностите на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на IRT, са независими от извадките, въз основа на които са получени, и поради това тези параметри са стабилни и инвариантни. Стабилността на параметрите също бе обоснована по-горе в текста.

(2) По отношение на (а) взаимовръзките между индексите на въпросите, определени в рамките на СТТ, и (б) взаимовръзките между параметрите, определена в рамките на IRT, в едно и също условие, т.е. при една и съща извадка от и. л.

(а) Допускаме, че между стойностите на индексите на трудност (p) и на дискриминативна сила (D и r_{bis}), определени в рамките на СТТ върху една и съща извадка, съществува взаимовръзка от нелинеен характер. Допускането се основава на теорията на данните на К. Кумбс, по-конкретно на модела „Данни единичен стимул” (Coombs, 1964). Ако един въпрос има екстремно висока/ ниска трудност, съгласно теоретичния модел той доминира над/ е доминиран от по-голяма част от индивидите. И в двата случая този въпрос би се характеризирал с екстремно ниски стойности на индекса на дискриминативна сила. Може да се предположи, че максималните си стойности този индекс би получил при въпроси със средна трудност, позиционирани в средата на скалата на съответния признак, подложен на измерване.

(б) Допускаме, че между стойностите на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на IRT, не съществуват взаимовръзки от корелационен тип или функционален тип. Това следва от вероятностния подход на тяхното оценяване, което предполага, че дадена характеристична крива може да заеме различна позиция на скалата на способностите Θ и същевременно да има различен (произволен) наклон или долна асимптота, в границите на изменение на съответния параметър.

(3) По отношение на съгласуваността между съответстващите си индекси и параметри, определени в едно и също условие, т.е. при една и съща извадка от и. л.

Допускаме, че между стойностите на оценките на трудността на въпросите p и b, както и между тяхната дискриминативна сила D и a и r_{bis} и a, няма съгласуваност. Това следва от теоретичната вариативност, нестабилност на индексите, определени в рам-

ките на CTT, и тяхната инвариантност в рамките на IRT. Ако даден индекс варира в допустимите граници на неговото изменение, а съответният параметър е стабилен, не би могло да се очаква да има съгласуване между техните стойности.

1.3. Задачи на изследването

Във връзка с формулираната по-горе цели и допускания, следва да бъдат изпълнени следните задачи:

(1) Да се идентифицира съвкупност/ съвкупности от тестови въпроси, които са изпълнявани от различни (поне две) групи от и. л.

(2) Да се идентифицират тестовите варианти, в които са включени тези въпроси.

(3) Да се формират двойки от тестови варианти, в които са включени едни и същи тестови въпроси.

(4) Да се изчислят индексите на трудност (p) и дискриминативна сила (D), определени в рамките на CTT, на въпросите от съответната двойка варианти

(5) Да се изчислят параметрите на трудност (b), дискриминативна сила (a) и налучкване на правилния отговор (c), определени в рамките на IRT, на въпросите от съответната двойка варианти.

(6) Да се приложат адекватни статистически методи за анализ на стабилността/ нестабилността на индексите на въпросите по CTT и съответните параметри по IRT.

(7) Да се приложат адекватни статистически методи за анализ на взаимовръзките между разноименните индекси и параметри, определени в рамките на съответната теория.

(8) Да се приложат адекватни статистически методи за анализ на съгласуваността между едноименните индекси и параметри.

1.4. Методология

1.4.1. Дизайн

1.4.1.1. Променливи величини и статистически методи

Планираното изследване принадлежи към категорията на корелационните изследвания. Този тип изследвания се прилагат често при анализите в сферата на психологическите и образователни измервания - в случаите, в които условията за провеждане на експериментално изследване са неизпълними или неподходящи (Gribbons & Herman, 1997). Най-общо корелационните изследвания се фокусират върху оценката на силата и типа на взаимовръзката между две или повече променливи. Степента на взаимовръзка между величините, които представляват интерес, се определя чрез някой от коефициентите на корелация, подходящ за съответния тип данни. Разкриването типа и структурата на тази взаимовръзка се счита за задължителна основа на по-

нататъшния задълбочен анализ на данните (Калинов, 2010).

За да се изследва стабилността/ нестабилността на една статистика на тестовите въпроси, нейната независимост/ зависимост от конкретната извадка от и. л., е необходимо тази статистика да бъде наблюдавана поне двукратно, при едни и същи тестови въпроси, които са попаднали в различни тестови варианти, използвани в различни изпитни сесии. В този смисъл по-нататък в текста, когато говорим за тестови варианти, ще имаме предвид не толкова техните поредни номера, колкото обстоятелството, че те са използвани в различни тестови сесии, на които са се явили различни групи от кандидат-студенти.

Двукратното наблюдение на дадена статистика, направено върху всеки един от множество въпроси, използвани в два различни тестови варианта, предполага формирането на два вектора със стойности на тази статистика. Като мярка за степента на устойчивост, стабилност на статистиките на въпросите, на тяхната независимост от условията, при които са оценени, ще бъде използван коефициент на корелация, подходящ за типа на скалата, която формира съответната статистика. Параметрите на въпросите по IRT образуват интервална скала и поради това към тях ще бъде приложен Пиърсъновия коефициент на корелация между стойностите на съответната статистика, изчислени въз основа на резултатите от първата и втората извадка от и. л. За изследване на стабилността на индексите на въпросите в рамките на CTT, за които се предполага, че образува рангова скала, ще бъде приложен непараметричният коефициент на рангова корелация R на Спиърмън. Този коефициент се разглежда като специален случай на линейния коефициент на корелация на Пиърсън и предполага, че съответните променливи са измерени поне в порядкова скала (Калинов, 2010). Ще бъде направен опит за изследване на стабилността на тези индекси и чрез параметричен коефициент на корелация, като особено внимание ще бъде отделено на трудността на въпросите (p).

Разбира се, при това изследване статистическата функционалност на корелационното изследване ще бъде приложена в по-редуциран вид. Тук не може да се говори, в строгия смисъл на думата, за функционална взаимовръзка между двете променливи, т.е. между двете оценки на съответната статистика. Психометричният смисъл на корелационния анализ е той да бъде в услуга на изследването на стабилността/ неизменяемостта/ съгласуваността на една или друга оценка на качеството на тестовите въпроси. Неговото приложение тук ще бъде ограничено до установяване на това дали има и каква е степента на съгласуваност между стойностите, получени при двете оценки. Поради качеството му, в което се използва в това изследване, корелационният коефициент следва да бъде разглеждан като „коефициент на стабилност" на статистиките на въпросите.

Имайки предвид границите на изменение на коефициентите на корелация, техните високи стойности, близки до 1.00, следва да бъдат интерпретирани като свиде-

телство за устойчивост на съответната статистика, за нейната независимост от условията на извадката. По-надолу ще бъдат фиксирани конкретни прагови стойности.

За верифициране на останалите допускания също ще бъдат приложени съответстващи на типа на променливите корелационни методи.

Поради характера на настоящото изследване в качеството си на променливи величини в него са включени следните статистики на въпросите:

(1) индекс на трудност (p)

(2) индекс на дискриминативна сила (D)

(3) бисериален коефициент на корелация (r_{bis}), определени в рамките на CTT

(4) параметър на дискриминативна сила (a)

(5) параметър на трудност (b)

(6) параметър на налучкване на правилния отговор (c), определени в рамките на IRT

1.4.1.2. Критерии за оценка на стабилността

Известно е, че коефициентът на Пиърсън е мярка за линейна корелация и може да достигне високи равнища не само тогава, когато стойностите на двете променливи по всеки обект (тестов въпрос) са равни или близки, но и тогава, когато стойностите на едната променлива са системно по-високи/ по-ниски от тези на другата. В този случай разпределенията на статистиките могат да имат различни средни и, възможно, различни стандартни отклонения. Поради това корелационният коефициент на стабилност може да даде информация доколко наблюденията при второто измерване са възпроизвели относителните позиции на наблюденията в първото.

При променливите, измерени в рангова скала, може да се наблюдава същият ефект, ако стойностите на отделните скали принадлежат към различни подмножества от числовата система с отношения.

Статистиките на въпросите се използват като мярка за измерителните качества на съответния въпрос. Те са обект на внимателен анализ при процедурата на позитивен/ негативен подбор на въпросите след пилотното тестиране, за формиране на окончателния вариант на теста. Поради това увереността (ако има основания за такава увереност), че статистиките запазват относителните си позиции, е може би недостатъчна за признаване на тяхната стабилност. Например, ако бъде установена висока корелация между индексите за трудност на въпросите (p), това би могло да означава, че ако при един пилотен тест (първа извадка) даден въпрос j_1 има стойност $p_{\text{j1}} = 0.43$ и бъде оценен като качествен, то при един актуален тестов вариант (втора извадка) същият въпрос j_1 може да придобие стойност $p_{\text{j1}} = 0.03$ и да влоши общото качество на теста.

Ето защо като втори, съпътстващ метод за оценка на стабилността на тестовите статистики ще приемем съотношението между централните им тенденции при първото

и второто измерване. Подходящ метод за оценка на такъв тип съотношения при променливи, измерени в интервална скала, е дисперсионния анализ с повторни измервания (*Repeated measures ANOVA*) с една независима променлива – вариант на теста (т.е. условие, при което е получена съответната статистика, което съответства на извадката от и. л., от резултатите на които е изчислена тази статистика) с две равнища, които условно ще обозначим като първа и втора оценка на съответната статистика. В този случай ще бъдат формулирани серия от нулеви хипотези за всяко сравнение с общ вид:

$$H_0 : \mu_1 = \mu_2$$

където: μ_1 и μ_2 – математически очаквания на стойностите на съответния индекс/ параметър при първата и втората оценка

При неметричните скали като тази на индекса на дискриминативна сила по CTT, за който може да се предполага, че образува рангова скала, ще бъде приложен, подобно на корелационния анализ, по-подходящ статистически метод. Това е Знаково-ранговия тест на Уилкоксън за зависими извадки (*Wilcoxon matched pairs test*), който е непараметричен аналог на *t*-теста на Стюдънт за зависими извадки или на ANOVA с повторни измервания. Нулевата хипотеза, която подлежи на проверка е, че променливата, образувана от разликите ($d = x - y$) между всяка двойка стойности (x, y) има нулева медиана.

В по-схематична форма нулевата хипотеза може да се представи като равенство на медианите на разпределенията на двете изходни променливи:

$$H_0 : Me_x = Me_y$$

Съобразно двата подхода за изследване на стабилността на тестовите статистики, като критерий за оценка на тяхната вариативност/ инвариантност ще приемем следния конюнктивен модел, който включва две условия, които следва да бъдат удовлетворени едновременно:

(1) Стойности на коефициента на стабилност, равни на 0.70 или по-високи, ще бъдат интерпретирани като свидетелство за инвариантност на съответния индекс/ параметър.

Макар че корелационният анализ е може би най-широко използваният статистически метод в областта на психологическите изследвания, могат да бъдат приведени множество примери за различни интерпретации (оценки) на големината на получените коефициенти, което означава, че по този въпрос няма конвенция. Често цитирани са критериите (по-скоро – практически правила), предложени от Дж. Коен за оценка на големината на корелационния коефициент като мярка за размера (силата) на ефекта. Дж. Коен предлага коефициенти със стойност около 0.10 да се разглеждат като ниски, тези около 0.30 – като средни/ умерени и тези със стойности около 0.50 – като високи

(Cohen, 1988, стр. 77–81). Авторът базира тези граници на обичайните, най-често наблюдавани стойности в при изследвания в областта на поведенческите науки и образователните измервания. Дж. Хемфил прави интересен опит за разпростре тази класификация върху резултатите от корелационни изследвания в тези области, в които корелационните коефициенти се използват по обичайното им предназначение (Hemphill, 2003). Авторът анализира 380 мета-аналитични изследвания в областта на психологическите измервания и терапии, извличайки докладваните в тях корелационни коефициенти и размери на ефекта (последните също са трансформирани в корелационни коефициенти), сортира ги по възходящ ред на абсолютните стойности, след което ги разпределя в 3 последователни групи с приблизително еднакъв обем. Коефициентите в първата група като цяло са по-ниски от 0.20, във втората варират от 0.20 до 0.30, а в третата са над 0.30. Авторът предлага точно тези стойности като гранични, още повече, е намира много сходства между полученото от него емпирично разпределение на корелационните коефициенти и тези, получени при други изследвания, цитирани от него. Интересно е да се отбележи, че максималната стойност на коефициентите на корелация в горната третина, наблюдавани в изследвания в областта на психологическите измервания, е 0.78.

Макар че изглеждат твърде либерални, референтните стойности на Дж. Коен са по-скоро строги. Например стойността на $r = 0.50$ за голям размер на ефекта съответства на 89-тия процентил от разпределението на коефициентите на корелация в областта на психологическите измервания. Това означава, че 89% от получените корелационни коефициенти имат стойности, равни или по-малки от 0.50 (ibid.)

Въпреки това смятаме, че праговата стойност на коефициента на стабилност следва да бъде много по-консервативна. Приемаме праговата стойност от 0.70 по подобие на някои други мерки, основани на корелацията, като коефициента на стабилност при оценката на надеждността чрез повторно използване на един и същи тест или коефициента на надеждност на Кронбах.

(2) Като свидетелство за инвариантност на съответния индекс/ параметър ще бъдат разглеждани и случаите на потвърдена нулевата хипотеза за липса на разлика между средните стойности (ANOVA с повторни измервания) или медианите (при теста на Уилкоксън) при поне среден/ умерен размер на ефекта.

За оценка на размера на ефекта в случаите на повторни измервания се използва коефициентът на частна корелация ета на квадрат (*partial eta-squared*), който отразява дела на вариацията на ефекта и на грешката в зависимите променливи, която може да бъде обяснена с въздействието на съответния фактор. За разлика от коефициента ета на квадрат, при коефициента на частна корелация няма общоприети норми за оценяване на размера на ефекта. Дж. Коен предлага практически правила, приложими главно към η^2, но които могат да се използват и при коефициента на частна корелация в случаите на еднофакторен дизайн: малък/ слаб – 0.01; среден/ умерен –

0.059; голям – 0.138 (Cohen, 1988).

1.4.1.3. Проблемът с типа на скалата на индекса на трудност (p)

Както е известно, съгласно СТТ индексът на трудност да въпросите се определя като съотношение между броя на правилните отговори към общия брой на отговорите, т. е. като относителен дял на правилните отговори, който може да бъде изразен чрез стойности в интервала $0.00 < p < 1.00$ или в проценти. Процентната скала на трудността на въпросите обаче не е интервална, а рангова и поради това използването на стандартния Пиърсънов коефициент на корелация като мярка на стабилността на този индикатор изглежда проблематично.

В своите влиятелните теоретични разработки С. Стивънс предлага не само йерархична система на измервателните скали, която се използва активно както в теоретични, така и в емпирични изследвания в автентичния си или в модифициран вид (Stevens, 1939, 1946; Стивънс, 1960). Той обвързва характеристиките на типовете скали с определени (допустими) статистики, които могат да се приложат над получените числови стойности. Освен спорния въпрос за номиналната скала, дискусионен все още е въпросът за разграничаването на ординалния и интервалния тип скали и съответно прилагането на непараметрични и параметрични статистики. За разлика от непараметричните, параметричните статистики (t и F тестове, Пиърсънови корелации, факторен и дисперсионен анализ и др.) изискват оценката поне на един параметър на разпределението на генералната съвкупност, което следва да е нормално.

Решителното обвързване на двете скали със съответните статистики и категоричното противопоставяне на използването на параметрични статистики върху ординални данни се дължи на С. Сийгъл (Siegel, 1956, по Gardner, 1975). Авторът определя 5 допускания при използването на параметрични статистики, четвъртото от които е, че променливите следва да бъдат измерени „поне в интервална скала, за да бъде възможно използването на аритметични операции" (Siegel, 1956, стр. 19, по Gardner, 1975, стр. 48). Дж. Гайто отбелязва, че това допускане е най-сериозната пречка за използване на параметрични техники, тъй като в много психологически изследвания се борави със суб-интервални данни (Gaito, 1960, по Gardner, 1975).

Много изследователи приемат и следват стриктно тази концепция (виж Гласс и Стэнли, 1976). Други обаче я подлагат не само на съмнение, но и на основателна критика. Разликата между ординалната и интервалната скала не е черно-бяла, смята П. Гарднър. Авторът поддържа тезата, че много от получените емпирични скали заемат сивата зона между тях (Gardner, 1975). В своя обстоен преглед на развитието на полемиката по тази тема той представя гледищата на редица автори, които оспорват „забраната", наложена от С. Сийгъл. Още Р. Абелсън и Дж. Тюки изказват мнението, че недостигът на метрична информация не означава непременно наличието на рангова информация, а обикновено нещо повече от нея (Abelson & Tukey, 1959, по Gardner,

1975). Дж. Гайто отбелязва, че разликата между двете скали (и техните допустими статистики) изобщо не е рязко очертана и че едни и същи данни могат да имат свойствата на две или повече скали (Gaito, 1960, по Gardner, 1975).

Аргументите срещу валидността на допускане 4. на С. Сийгъл са два типа: базирани на анализ на математическите характеристики на параметричните статистики и базирани на емпирични доказателства.

Аргументи от първия тип дават Ф. Лорд, Н. Андерсън, О. Кемпторн, Дж. Гайто, К. МакНемар и др., които са убедени, че нивото на значимост на параметричните тестове и валидността на статистическия извод не зависят от типа на измервателната скала (Gardner, 1975). Б. Бейкър, К. Хардик и Л. Петринович обобщават, че „формалните характеристики на измервателните скали като такива, не трябва да влияят върху избора на статистики" (Baker, Hardyck & Petrinovich, 1966, стр. 292, по Gardner, 1975, стр. 50).

Аргументи от втория тип представят С. Лейбовиц и Б. Бейкър, К. Хардик и Л. Петринович, които изучават ефекта от трансформацията на метричните свойства на скалата върху равнищата на значимост на различни статистики (Baker, Hardyck & Petrinovich, 1966; Labovitz, 1967, по Gardner, 1975). Авторите провеждат изследвания, в които прилагат различни линейни и нелинейни трансформации на променливите, след което прилагат различни параметрични статистики както върху изходните, така и върху резултативните променливи . Тяхното заключение е, че статистическите изводи почти не се влияят от типа на скалата.

П. Гарднър заключава, че „…поради устойчивостта (*robustness*) на параметричните техники, третирането на ординалните данни като интервални не би довело до погрешни заключения." (Gardner, 1975, стр. 51). Още по-категоричен е К. МакНемар, който, задавайки въпроса дали параметричните статистики ще следват своите теоретични извадкови разпределения, ако данните не са интервални, отговаря „…категорично да при условие, че разпределението на данните не се отклонява значително от нормалното." (McNemar, 1969, стр. 431, по Gardner, 1975, стр. 52).

При все това редица автори препоръчват, при съмнения за чувствително неравенство на интервалите в ординалната скала, данните да бъдат подложени на трансформация (Gardner, 1975; Taylor, 1985; Aiken, 1988; Micceri, 1989; Weiner et al., 2003). При трансформацията се променя скалата на измерване, което води до формиране на нова променлива, математически еквивалентна на изходната, но с нормално разпределение.

1.4.2. Процедура за подбор на въпросите

Определянето на въпросите, отговарящи на посочените по-горе условия и подходящи за включване в емпиричното изследване, бе извършено с любезното съдействие на Центъра по оценяване към НБУ, който осигури достъп до резултатите от изпитите (до суровите данни) от наличните тестови варианти на ТОП от 1998 г. до 2008 г.

Селекцията на въпросите бе извършена при спазване на изискванията на Центъра за конфиденциалност на информацията.

Първата задача при провеждане на емпиричното изследване бе да се идентифицират отделни въпроси (или групи от въпроси), които попадат в няколко (поне два) варианта на ТОП, използвани в различни тестови сесии.

Процедурата за подбор на въпроси, които отговарят на това условие, стартира с проучване на „паспортите" на отделните въпроси. Това са архивни записи, в които се съхранява и поддържа детайлна информация за тяхното администриране, включително и за тестовите варианти и изпитните сесии, в които са били използвани. След като бе установено наличието на такива въпроси, усилията бяха фокусирани върху това да се определи поне една, а при възможност – няколко по-големи групи от въпроси, всяка от които да е използвана при конструирането на два различни варианта, предназначени за различни изпитни сесии. След продължително проучване, в хода на което бяха отхвърлени множество въпроси, използвани в два или повече тестови варианта, но със слабо сечение помежду им, се откроиха три двойки от тестови варианти, в които се наблюдава почти пълно съответствие между въпросите във всички раздели на ТОП.

Справка за избраните (двойки) тестови варианти е представена в следващата таблица. При първата и втората двойка броят на общите въпроси е 99 (99.00% от всички), а при третата – 100 (100%).

Таблица 26. Данни за тестовите варианти на ТОП, използвани в емпиричното изследване

Първа оценка				Втора оценка		
Номер на двойка варианти	**Вариант на ТОП**	**Дата на изпитна сесия**	**Брой изпитани**	**Кореспондира с вариант на ТОП**	**Дата на изпитна сесия**	**Брой изпитани**
1	92	20.04.2003	636	128	27.02.2005	543
2	96	18.04.2004	652	132	17.04.2005	638
3	110	27.06.2003	865	146	24.07.2005	454

Информацията в таблицата показва, че тестовите варианти са използвани в различни изпитни сесии, при отделните двойки – в различни календарни години, с различен брой кандидат-студенти. Всичко това дава възможност, съгласно дизайна на изследването, за двукратна оценка на всяка статистика върху отделна, независима извадка.

1.4.3. Процедура за психометричен анализ на въпросите

След селектирането на двойките тестови варианти, суровите резултати от всеки вариант бяха обработени, по съответния ключ за правилните отговори, с програмния продукт Iteman, който е част от Item and Test Analysis Package, разработен от компани-

ята Assessment Systems Corp. (Assessment Systems Corporation, 1997). Iteman е специализирана професионална програма за анализ на тестовите данни по Класическата тестова теория. Анализът бе извършен на ниво „тест", т. е. при разглеждането на съответния тестова вариант като единна скала, включваща 100 въпроса. Сред различните резултати от приложението на алгоритмите на Класическата теория важни за настоящия анализ са числовите стойности на индексите на трудност (p) и дискриминативна сила (D) на въпросите от всеки тестов вариант.

След това суровите резултати от всеки тестов вариант бяха подложени и на процедура за калибриране, т. е. за оценка на параметрите на тестовите въпроси съгласно трипараметричния модел на Теорията за отговор на тестов въпрос. Обработката на данните беше направена със софтуерния продукт Xcalibre, който е част от модула за анализ на айтеми и тестове от психометричния софтуер MicroCAT™ Testing System (ibid.) Програмата е специализирана за извършване на анализ по дву- и трипараметричния логистичен модел на IRT. За целите на тази част от изследването бе приложен трипараметричният модел, в рамките на който бяха изчислени стойностите на параметрите дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c).

За калибриране на параметрите се прилага метода на маргиналното правдоподобие (*Marginal maximum-likelihood, MML*). Този метод се състои в определянето на такива оценки на неизвестните параметри на въпросите и на индивида, които да максимизират съответните функции на правдоподобието. Чрез този метод се получават асимптотично ефективни оценки на параметрите.

В алгоритъма на Xcalibre методът *MML* за калибриране на параметрите на тестовите въпроси се реализира на няколко стъпки.

(1) Начална фаза, в която се прави предварителна оценка на параметрите, основана на техните индекси по CTT (трудност p, бисериален коефициент на корелация r_{bis} в качеството му на дискриминативен индекс и на налучкване на правилния отговор като реципрочна стойност на броя на алтернативните отговори). Стойностите на класическите индекси се трансформират по определен алгоритъм в първоначални оценки на параметрите a, b и c.

(2) Прилагане на алгоритъма *EM* (*Expectation - Maximization*), при който като начин за оценка на параметрите се използва максимизирането на функцията на максималното правдоподобие. Сам по себе си този алгоритъм е циклична двустъпкова процедура, която при стъпка „*E*" цели да се определи очаквания брой на лицата в популацията, разпределени между 15 предварително фиксирани точки на континуума Θ (в интервала от -3.5 до 3.5, през 0.50), и дела на онези лица във всяка група, които биха отговорили правилно на съответния въпрос. На стъпка „*M*", която е итеративна, се прави оценка на параметрите на всеки въпрос до удовлетворяване на предварително определен критерий. Циклите *EM* за всички въпроси продължават до тогава, докато не бъде постигнат фиксирания критерий за толерантност и при най-„упорития" въпрос. Този

критерий представлява сумата от абсолютните стойности на измененията във всички параметри на даден въпрос при даден цикъл, в сравнение с предходния, и тази сума не трябва да надхвърля 0.05.

При оценката на параметрите се прилага Байесовия подход, съгласно който се предполага, че не само оценките, но и самите параметри са случайни величини, които имат някакво вероятностно разпределение. Плътността на разпределението на параметъра трябва да бъде известна преди да се направи неговата оценка, т. е. необходимо е да се определи някакво априорно разпределение на вероятностите. За да се подпомогне процеса на оценяване, за всеки параметър на въпросите в алгоритъма на програмата са определени различни априорни разпределения, с фиксирана средна стойност и стандартно отклонение.

1.4.4. Трансформиране на стойностите на индекса на трудност (p) в интервална скала

Една от обичайните процедури е нормализиране на данните, което се изразява в трансформирането на отделните стойности на изходното разпределение в z-единици на нормираното (стандартно) нормално разпределение. Л. Айкен отбелязва, че за разлика от ординалното измерване, „стандартизираните оценки представят измерването в интервална скала" (Aiken, 1988, стр. 87). Същата техника се препоръчва и използва от редица други изследователи (Fan, 1998; Анастази и Урбина, 2001).

Макар че, както бе отбелязано по-горе, една суб-интервална скала би могла да се третира като интервална, ние направихме допускането, че е възможно в скалата на трудността p да се наблюдава чувствително неравенство на интервалите и поради това е необходимо нейното нормализиране. За трансформацията на суровите стойности на индекса на трудността p в интервална скала бе приложена следната двустъпкова процедура.

(1) Преобразуване на стойностите на индекса p в процентили по формулата $P_i = (1 - p_i)$.

(2) Преобразуване на получените стойности в z-единици на стандартното нормално разпределение.

Тази процедура бе приложена към суровите стойности на индекса p, изчислени по съответния алгоритъм на СТТ за всички тестови варианти, обхванати в анализа. Тестовите варианти се разглеждат като единна скала, състояща се от 100 въпроса.

Първата стъпка се извършва поради това, че стойността на p съответства на относителния дял на лицата, които доминират над съответния въпрос съгласно теоретичния модел „Данни единичен стимул" на К. Кумбс (Coombs, 1964). Чрез това преобразуване се определя делът на лицата, които са доминирани от съответния въпрос, т.е. чиито идеални точки се намират наляво от неговата точка.

На втората стъпка, от статистическа таблица със стойностите на стандартното

нормално разпределение, по получените процентилни стойности, които съответстват на частта от лицето на повърхнината под кривата на стандартното нормално разпределение от -∞ до дадената точка, се определя съответната z-стойност (Анастази и Урбина, 2001; Калинов, 2010). Преобразувани по този начин, стойностите на индекса p формират нормална крива и се изразяват в единиците на нормираното нормално разпределение със средна стойност M = 0.00 и стандартно отклонение SD = 1.00.

При анализа на стабилността на този индекс са използвани стандартизираните z-стойности, което позволява прилагането както на Пиърсъновия коефициент на корелация, така и на дисперсионния анализ с повторни измервания.

2. Резултати

2.1. Оценка на адекватността на 1-, 2- и 3-параметричен модел на IRT

Преди да се проведат планираните съпоставителни изследвания на „очакваното" поведение на статистиките на въпросите, е необходимо да се провери дали планираното съпоставително изследване на три параметъра в рамките на IRT е обосновано, т. е. да се установи кой от моделите на IRT, по отношение на броя на параметрите, е най-адекватен на тестовите данни. Оценката на годността на трите модела е направена чрез програмите Rascal и Xcalibre, които са част от модула за анализ на айтеми и тестове от психометричния софтуер MicroCAT™ Testing System (Assessment Systems Corporation, 1997).

При еднопараметричния модел на Раш статистиката, която е използвана за оценка на (не)съответствието (*lack-of-fit*) на всеки въпрос с модела, е хи-квадрат на Пиърсън. За изчисляване на тестовата статистика, и. л. се разделят на последователни категории според оценките на личностовия им параметър. Броят на групите е максимум 20 и тази стойност се използва за определяне на степените на свобода на тестовата статистика. При конкретните анализи бе прието ниво на значимост α = 0.05. При това условие критичната стойност на χ^2 = 30.144, при 19 степени на свобода.

При дву- и три-параметричния модел като мярка за точността на оценките на параметрите, определени съгласно тествания модел, т.е. на неговата годност, се използват стандартизираните остатъци (*standardized residuals*). Стойността на даден стандартизиран остатък е индикатор за степента, в която статистиките на съответния айтем се съгласуват с очакванията, произтичащи от прилагания дву- или трипараметричен модел. Стандартизираните остатъци са разпределени нормално, а като критична стойност е избрана 2.00. Стойности, по-високи от тази, съответстват на тест за значимост за грешка от I род с приблизително стойност на α = 0.05

Като основна мярка за неадекватността на съответния модел е приет броят на въпросите, чиято тестова статистика попада в критичната област. В следващата таб-

лица са представени резултатите от прилагането на тестовете за адекватност на трите модела на IRT върху данните от тестовете, включени в анализа.

Таблица 27. Брой на въпросите, които не съответстват на модела

Вариант на ТОП	Модел на IRT		
	Еднопараметричен (b)	Двупараметричен (a, b)	Трипараметричен (a, b, c)
вар. 92	24	29	6
вар.96	25	24	6
вар.110	29	19	6
вар.128	30	25	7
вар.132	27	27	4
вар.146	17	29	10

Броят на въпросите, които не съответстват на тестваните едно-, дву- и трипараметрични модели, е различен при отделните тестови варианти. Като цяло той е значително по-голям при едно- и двупараметричния модели, при които варира съответно от 17 до 30 въпроса със средна стойност 25.33 и от 19 до 29 въпроса със средна стойност 25.50. Броят на въпросите, които не съответстват на трипараметричния модел, е между 4 и 10, със средна стойност 6.50. Несъмнено моделът, който най-добре описва данните, е трипараметричният. Разликата между него и двупараметричния модел е във включването на параметъра за налучкване на коректния отговор c, за който следва да предположим, че играе съществена роля като елемент от функционалната връзка между латентното пространство и начина, по който и. л. отговарят на въпросите в теста.

2.2. Описателни статистики на зависимите променливи

Описателните статистики представят обобщена информация за разпределенията на зависимите променливи (индекси в рамките на Класическата тестова теория и параметри – на Теорията за отговор на тестов въпрос). Те биха могли да дадат полезна информация за общото равнище на стойностите на съответната статистика, на тяхната хомогенност в рамките на анализираните тестови варианти, както и да послужат за първоначална оценка на качеството на тестовете като цяло. Всяка от статистики на въпросите са изчислени чрез алгоритмите на съответната теоретична рамка на описаните в теоретичната част тестови теории.

2.2.1. Характеристики на въпросите съгласно Класическата тестова теория

Ще започнем представянето на тестовите статистики с индексите на дискриминативна сила на въпросите, определени чрез алгоритмите на Класическата теория. Следващата таблица съдържа основните описателни статистики на неговото разпре-

деление в шестте анализирани тестови варианта.

Таблица 28. Описателни статистики на дискриминативния индекс (*D*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.201	-0.040	0.550	0.017	0.129
	вар. 128	99	0.203	-0.110	0.520	0.019	0.137
2.	вар. 96	100	0.208	-0.050	0.530	0.015	0.124
	вар. 132	99	0.216	-0.030	0.560	0.018	0.133
3.	вар. 146	100	0.193	-0.060	0.450	0.013	0.115
	вар. 110	100	0.206	-0.090	0.490	0.015	0.123

Първото нещо, което прави силно впечатление, са относително ниските средни стойности на този индекс, които се наблюдават във всички тестови варианти. Максималната средна в горната таблица е тази при вариант 132 (0.216), а минималната - при вариант 92 (0.201). Сравнително ниските средни стойности на индекса на дискриминативна сила, които при различните тестови варианти са устойчиво стабилизирани в диапазона от 0.20 до 0.22, са свидетелство, че тестовите варианти като инструменти на измерване не разграничават добре лицата с по-ниски от тези с по-високи способности. Това обстоятелство е тревожно и от прагматична гледна точка, тъй като, съгласно едно установено практическо правило, минималната стойност на индекса D, която се приема за долна граница на приемливост, е точно 0.20 (Ebel, 1954).

Макар че всички варианти съдържат въпроси с негативни стойности на този индекс, добрата новина е, че са малко на брой (около 4% – 6% от въпросите във всеки вариант), а и отклоненията наляво от нулевата стойност са слаби. По-съществени са те при вариант 128, който включва въпроси с най-ниска стойност на този индекс (-0.110). От друга страна, най-високите стойности на този индекс са в диапазона 0.45 до 0.55, което е далеч от горната граница на неговото изменение.

Сходството между минималните и максималните стойности на индекса в отделните тестови варианти намира отражение в приблизително еднаквите оценки на тяхното разсейване. Стойностите на стандартните отклонения са в диапазона от 0.115 при вариант 146 до 0.137 при вариант 128. Независимо от това, че стойностите на този индекс едва прекрачват зоната отвъд нулевата стойност и не надхвърлят умерените равнища от 0.50 -0.60, степента на тяхното разсейване едва ли може да бъде пренебрегната. С други думи, дискриминативната сила на въпросите е вариативен признак, по който те се различават и който следва да бъде включен при тяхното моделиране в IRT.

Бисериалният коефициент на корелация е втора, статистическа мярка за разграничителната способност на въпросите. По същество двата индекса (заедно с „класическия” индекс) са източник на една и съща информация,още повече, че имат едни и

същи граници на изменение. Практиката показва, че бисериалният коефициент е по-консервативен от класическия. Това се потвърждава и от данните в следващата таблица.

Таблица 29. Описателни статистики на бисериалния коефициент на корелация (r_{bis})

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.199	-0.270	0.500	0.021	0.144
	вар. 128	99	0.207	-0.240	0.560	0.026	0.162
2.	вар. 96	100	0.233	-0.210	0.600	0.023	0.150
	вар. 132	99	0.233	-0.070	0.520	0.021	0.145
3.	вар. 146	100	0.188	-0.180	0.610	0.017	0.131
	вар. 110	100	0.198	-0.220	0.520	0.017	0.130

Средната стойност на този индекс се мени в сравнително тесни граници, като максималната стойност от 0.233 се наблюдава при два тестови варианта (96 и 132), минималната му стойност е 0.188 при вариант 146. Като цяло средните стойности на този статистически индекс на различителната сила са малко по-ниски от тези на класическия.

Забелязват се обаче значително по-ниските минимални стойности, които при всички тестови варианти имат отрицателен знак, за да достигнат до -0.270 при вариант 92. От друга страна, максималните стойности са малко по-високи в сравнение с тези от предходната таблица, достигащи до 0.610 при тестов вариант 146. Тази по-широк диапазон на изменение на бисериалния коефициент намира израз в по-високите стойности на стандартните отклонения в сравнение с тези при класическия индекс. Както би могло да се очаква, анализът на бисериалния коефициент като втора мярка на разграничителната способност на въпросите потвърждава направените по-горе оценки, че като цяло тестовите въпроси не различават добре лицата от контрастните групи и още, че тази характеристика на въпросите е вариативен признак, по който те се различават и който следва да бъде включен при тяхното моделиране в IRT.

Трудността на въпросите е характеристика, която влияе в най-висока степен на тестовите резултати, а и на трудността на теста като цяло. Данните от следващата таблица показват, че отделните тестови варианти се характеризират с приблизително еднаква средната трудност на въпросите, която се изменя в диапазона от 0.41 до 0.45. Поставена в контекста на теоретичните граничните стойности на този индекс, трудността на въпросите от различните варианти следва да се разглежда като превишаваща средното равнище.

За разлика от индексите на различителна сила, трудността варира в много широк диапазон. Минималните стойности при всички варианти се приближават към дол-

ната граница на изменение на този индекс, достигайки до 0.030 (при варианти 92 и 128), максималните – към горната граница на изменение, достигайки до 0.990 при вариант 110.

Таблица 30. Описателни статистики на индекса на трудност (*p*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.432	0.030	0.980	0.051	0.225
	вар. 128	99	0.417	0.030	0.980	0.050	0.224
2.	вар. 96	100	0.439	0.040	0.970	0.060	0.248
	вар. 132	99	0.449	0.070	0.970	0.058	0.242
3.	вар. 146	100	0.408	0.060	0.980	0.051	0.226
	вар. 110	100	0.409	0.060	0.990	0.052	0.227

Сходните средни, минимални и максимални стойности на този индекс намират отражение в относително високите (заради големия размах), но приблизително еднакви стойности на стандартните отклонения. Следователно, съдейки по тези основни характеристики на разпределенията на индексите на трудност, можем да заключим, че в отделните тестови сесии, проведени през различни години, трудността на вариантите се проявява като тяхна устойчива характеристика, която не се изменя съществено.

2.2.2. Характеристики на въпросите съгласно Теорията за отговор на тестов въпрос

Числовите стойности (оценките) на параметрите, изчислени въз основа на алгоритмите на трипараметричния модел на Теорията за отговор на тестов въпрос, са получени в рамките на съвършено различен психометричен модел. Поради това описателните статистики на техните разпределения не могат да бъдат пряко съпоставени, но ние ще очертаем някои паралели, основани на възможните (или типичните) граници на техните изменения, както и на наблюдаваното равнище на хомогенност р рамките на съответните тестови варианти.

В следващите няколко таблици са представени основните описателни статистики на разпределенията на параметрите на въпросите в различните тестови варианти, предмет на настоящото изследване.

Средните стойности на дискриминативния параметър (a) при различните варианти се движат в сравнително тесните граници между 0.507 (при варианти 92 и 146) и 0.594 (при вариант 132). Интересно е, че се наблюдава известна разлика в средните равнища на този параметър между двойките варианти, съставени от едни и същи въпроси. Без да надценяваме тези наблюдения (поради липса на данни за значимостта на тези разлики), ще отбележим, че те могат да бъдат сигнал за определена вариативност на стойностите на този параметър.

Таблица 31. Описателни статистики на дискриминативния параметър (*a*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	0.507	0.320	0.860	0.012	0.110
	вар. 128	99	0.554	0.340	0.850	0.013	0.114
2.	вар. 96	100	0.574	0.360	0.890	0.013	0.112
	вар. 132	99	0.594	0.370	0.860	0.013	0.114
3.	вар. 146	100	0.507	0.350	0.740	0.008	0.089
	вар. 110	100	0.525	0.300	0.900	0.015	0.124

Друга особеност, която следва да бъде посочена, е липсата на въпроси с негативни стойности по този параметър, за разлика от едноименния индекс, определен по СТТ. Минималните наблюдавани стойности на въпросите при всички тестови варианти са положителни, като най-ниската сред тях е 0.300 (при вариант 110), а най-високата достига 0.900 (при същия вариант).

Съпоставени с едноименните индекси по Класическата теория, дискриминативните параметри са малко по-хомогенни, с по-слаба вариация. При по-голяма част от анализираните тестови варианти стандартното отклонение варира около 0.11 - 0.12, при 0.12 – 0.14 за класическия индекс D и 0.13 - 0.16 за бисериалния коефициент на корелация r_{bis}. С други думи, дискриминативният параметър на въпросите варира в степен, която е съпоставима с тази на индексите от класическата теория и поради това включването му в моделирането на въпросите изглежда оправдано.

Подобно на данните за класическия индекс на трудност, средните стойности на едноименния параметър свидетелстват за това, че въпросите се отличават с повишена трудност (в границите на изменение от ± 3.00, наложени от използвания софтуер), която варира от 1.102 (при вариант 132) до 1.401 (при вариант 110). Тук също се забелязват различия между средната трудност в рамките на отделните двойки варианти, което би могло да бъде индикация за вариативност на този параметър

Таблица 32. Описателни статистики на параметъра трудност (*b*)

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Макс.	Дисперсия	Станд. откл.
1.	вар. 92	100	1.280	-3.000	3.000	2.948	1.717
	вар. 128	99	1.271	-3.000	3.000	2.638	1.624
2.	вар. 96	100	1.122	-3.000	3.000	3.398	1.843
	вар. 132	99	1.102	-3.000	3.000	3.250	1.803
3.	вар. 146	100	1.386	-3.000	3.000	2.871	1.694
	вар. 110	100	1.401	-3.000	3.000	2.760	1.661

Друго сходство с класическия индекс може да бъде открито в широките граници на изменение на стойностите на параметъра, които покриват целия посочен по-горе

интервал. Това намира отражение във високите стойности на стандартното отклонение, които обаче надвишават многократно тези при класическия индекс. Обяснение за тази разлика може да бъде намерено в разпределенията на стойностите на съответните статистики. Докато разпределенията на класическия индекс се приближават към нормалното (с положителна асиметрия), то разпределенията на съответния параметър са по-скоро L-образни, с отрицателна асиметрия и с ярко изразен таванен ефект. Източникът на високите стандартни отклонение е натрупването на високи честоти в десния край на разпределенията.

Статистиките на разпределенията на параметъра на склонността към налучкване на правилните отговори са интересни преди всичко с това, че съдържат свидетелства за жизнеността на тази характеристика на въпросите. Средните стойности на параметъра в различните тестови варианти са близки, около 0.17 – 0.19, което е малко под очакваната стойност от 0.20 (1/5, съобразно броя на дистракторите във всички въпроси.

Таблица 33. Описателни статистики на параметъра склонност към налучкване

Двойка тестове	Вариант	Брой въпроси	Средна ст-т	Мин.	Мак.с	Дисперсия	Станд. откл.
1.	вар. 92	100	0.176	0.090	0.210	0.000	0.021
	вар. 128	99	0.178	0.100	0.210	0.000	0.021
2.	вар. 96	100	0.188	0.100	0.260	0.001	0.026
	вар. 132	99	0.192	0.110	0.240	0.001	0.023
3.	вар. 146	100	0.169	0.110	0.190	0.000	0.016
	вар. 110	100	0.173	0.100	0.210	0.000	0.022

Интересно е, че максималните стойности надвишават слабо средните, което очертава друга особеност на разпределенията на този параметър. Те са силно асиметрични, скосени отдясно, с концентрация на високи честоти непосредствено около средната стойност и ниски, намаляващи честоти под стойности от 1.15 – 0.16. Тази особеност намира отражение и в статистиките на разсейването, които имат много ниски, приблизително еднакви стойности. Всичко това показва, че склонността към налучкване е устойчива характеристика на въпросите, която, заедно с тяхната дискриминативна сила, не трябва да бъде пренебрегвана при тяхното описание.

2.3. Рангова ли е скалата на трудността?

Както бе отбелязано по-горе в текста, трудността на въпросите (p) се изразява чрез дела на правилните отговори от общия брой на отговорите на даден въпрос, често изразявана в проценти. Следствие от този начин на определяне на стойностите на индекса p е, че скалата, която формират неговите стойности, е рангова. За решаване на този проблем стойностите на p бяха трансформирани в z-единици на стандартното

разпределение. Л. Айкен, А. Анастази и С. Урбина и други изследователи отбелязват, че стандартизираните по този начин стойности формират интервална скала (Aiken, 1988; Анастази и Урбина, 2001).

П. Супес и Дж. Зинес подхождат към въпроса за типа на измервателната скала по-формално. В труда си „Basic measurement theory", формулирайки втората основна теорема за единствеността измерването, авторите изтъкват, че „от математическа гледна точка определянето на типа измервателната скала, приложена към дадена емпирична система, се определя от начина, който позволява да се премине от една числова система към друга, ако те включват едни и същи отношения и са хомоморфни на една и съща емпирична система" (Суппес и Зинес, 1967, стр. 18).

Авторите предписват следните ограничения пред преобразуването на отделните скали (φ), което се отбелязва и като „допустимо" преобразуване на дадена скала. За интервалните скали допустимото преобразуване φ е положително и линейно, т. е. трябва да има едно положително действително число α и едно число β, което може да бъде положително, отрицателно или нула, за които, при всяко действително число x, $\varphi(x) = \alpha\, x + \beta$. При ранговите скали преобразуването φ е всяко монотонно преобразуване, което може да бъде нарастващо или намаляващо. При първия тип монотонно преобразуване за всяко x и y от неговата област $\varphi(x) < \varphi(y)$ само тогава, когато $x < y$. Функцията φ е монотонно намаляващо преобразуване, ако за всяко x и y от неговата област $\varphi(x) > \varphi(y)$ само тогава, когато $x < y$ (ibid.)

След трансформирането на суровите стойности на трудността p в z-единици на стандартното нормално разпределение, този индекс е представен чрез два типа скали, в две числови системи, които включват едни и същи отношения и са хомоморфни на една и съща емпирична система от тестови айтеми. Ако суровите стойности на индекса p действително образуват рангова скала, то взаимовръзката им със скалата на z-единиците следва да бъде монотонно, а не линейно преобразуване. За проверка на това предположение бяха изчислени коефициентите на Пиърсънова корелация между отделните двойки скали за всеки тестов вариант, представени в следващата таблица.

Таблица 34. Коефициенти на линейна корелация между суровите и стандартизираните стойности на индекса на трудност (*p)*

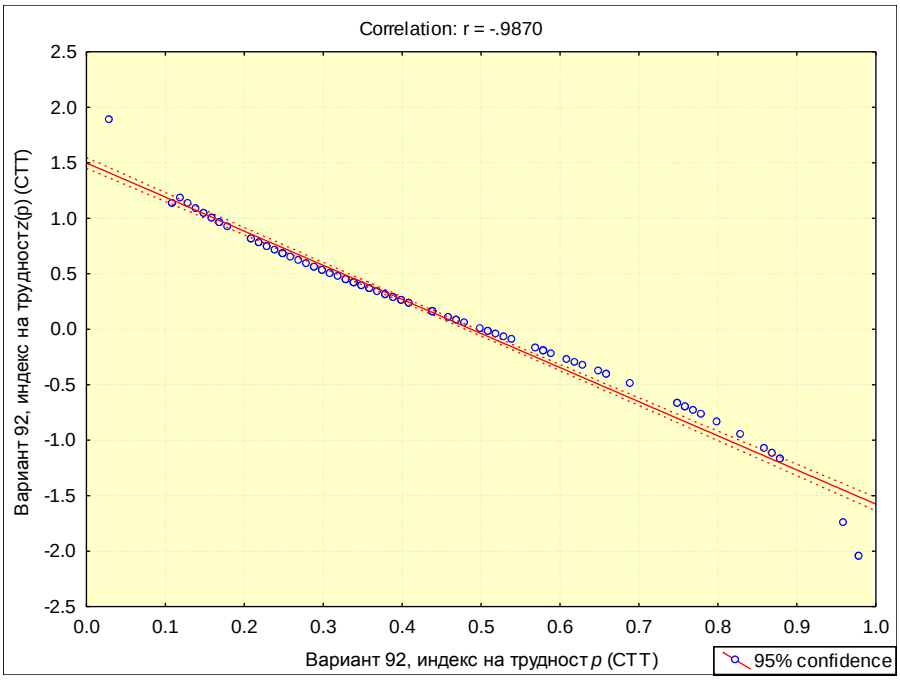
	вар. 92 *z*(p)	**вар. 128 *z*(p)**	**вар. 96 *z*(p)**	**вар. 132 *z*(p)**	**вар. 146 *z*(p)**	**вар. 110 *z*(p)**
вар. 092-p	*-0.987					
вар. 128-p		*-0.988				
вар. 096-p			*-0.992			
вар. 132-p				*-0.991		
вар. 146-p					*-0.992	
вар. 110-p						*-0.989

Забележка: Всички коефициенти на корелация са значими при p < 0.05

Наблюдаваните коефициенти на корелация при всички двойки p – $z(p)$ имат изключително високи стойности, около и над -0.99, които са значими при $p < 0.05$. При това тези коефициенти отразяват линейната корелация между стойностите на две скали, за които предполагаме, че са свързани с монотонна връзка. Ако това е така, коефициентите на корелация биха представлявали занижени оценки на силата на взаимовръзката между съответните две променливи (Калинов, 2010). Отрицателният знак се дължи, разбира се, на начина на трансформиране на суровите стойности на индекса p в z-единици. При наличието на такива високи стойности на линейните корелационни коефициенти между двете скали би могло да се каже, че преобразуването на p в z-единици е линейно и следователно скалата на трудността p, съгласно разгледаните по-горе допустими преобразувания, може да се третира като интервална скала.

За да потвърдим това мнение, би било полезно да разгледаме диаграмите на разсейването на стойностите в двумерното пространство, образувано от двете скали. Като илюстрация по-долу е представена диаграмата на стойностите на p и z за вариант 92.

Фигура 21. Връзка между суровите (*p*) и стандартизираните стойности *z*(*p*) на индекса на трудност при вариант 92



Строго погледнато, функцията, изразяваща връзката между суровите и стандартизираните стойности, следва да се приеме за монотонна (монотонно намаляваща). Сравнително ясно се откроява S-образната ѝ форма, особено в двата края на скалата на p.

От друга страна, забелязва се характерното за случаите на висока линейна корелация подреждане на точките на айтемите на или в непосредствена близост до регресионната линия, по-специално в интервала (0.10 – 0.90). Ще обърнем внимание и на

тесните граници на 95% доверителен интервал.

Изключение от (почти) линейната подредба правят, както беше отбелязано, само малък брой айтеми в двата края на скалата p с екстремни стойности на трудността, близки до 0.00 или до 1.00.

За да направим статистически обоснован извод за формата на връзката между двете оценки на p, ще проверим хипотезата за нулева стойност на регресионния коефициент b:

$$H_0 : b = 0.00$$

Таблица 35. Оценка на регресионния коефициент за вариант 92

	b	**Стандартна грешка на *b***	***B***	**Стандартна грешка на *B***	***t* (97)**	***p***
свободен член			1.499	0.025	60.96	0.000
p (вариант 92)	*-0.987	0.016	*-3.074	0.051	-60.87	0.000

Забележка: Маркираните коефициенти са значими при p < 0.05

Нулевата хипотеза е проверена чрез регресионен анализ, в който суровата оценка на трудността (p) играе ролята на независима променлива. Тази хипотеза може да бъде отхвърлена на ниво α = 0.05. Данните сочат не само за високата стойност на ъгловия коефициент, но и за ниската стандартна грешка на неговата оценка. Цялостната оценката на годността на модела, направена чрез ANOVA, е също висока – F(1, 98) = 3705.376, p = 0.000. Изравненият коефициент на детерминация R^2 квадрат има стойност 0.974. Той отразява онази част от дисперсията на зависимата променлива в популацията от въпроси, която може да бъде обяснена чрез построения линеен модел. Наблюдаваната стойност на този коефициент също е свидетелство за неговото високо качество.

Ще използваме и още един независим индекс за силата на взаимовръзката между двете променливи. Това е коефициентът ета (η), популярен и като корелационно отношение, който се използва предимно като мярка за оценка на нелинейната взаимовръзка между две променливи, макар че по същество той е мярка за всякакъв тип взаимовръзки (Гласс и Стэнли, 1976; Калинов, 2002). Корелационното отношение се определя като отношение на вътрешногруповата сума на квадратите към общата им сума:

$$\eta_{y,x} = \sqrt{1 - \frac{SS_{between}}{SS_{total}}} \qquad (64)$$

Стойността на този коефициент по данните от разгледания по-горе вариант 92 е η = 0.994, която е малко по-висока от стойността на съответния Пиърсънов корелационен коефициент. Тъй като диаграмата на разсейване от фигура 21 представя свидетелства за наличието на нелинеен компонент в съвместното поведение на двете про-

менливи, бихме могли да направим оценка на степента на нелинейност чрез разликата η^2 - r^2 (Калинов, 2010). Според данните от вариант 92, тази разлика е 0.994^2 – (-$0.987)^2$ = 0.013.

Изследване 3. Анализ на инвариантността на статистиките на тестовите въпроси

2.4. Стабилност на индексите на въпросите, определени в съответствие с Класическата тестова теория

Както беше отбелязано, важно качество на тестовите въпроси е стабилността на техните индекси. Ако техните стойности се изменят в зависимост от тестовата сесия, в която са използвани, съответно от извадката, въз основа на която са изчислени, това не само би подронило доверието в тези индекси като характеристики на въпросите, но би възпрепятствало и създаването на еквивалентни тестове.

Поради ранговия характер на тези индекси, като първа оценка на тяхната стабилност са използвани коефициентите на рангова корелация R на Спиърмън и, в допълнение, на линейна корелация r_{xy} на Пиърсън, приложени върху две серии от стойности на съответния индекс, получени от едни и същи съвкупности от тестови въпроси, използвани в две различни условия, описани по-горе. Ако при това съпоставяне се наблюдават високи равнища на корелация и по-конкретно – по-високи от приетата прагова стойност, то това би означавало, че съответната статистика е стабилна и не се влияе от извадката, въз основа на която е изчислена.

В допълнение, като втора мярка на стабилността на индексите е направена статистическа оценка на разликата между техните средни стойности или медиани. Ако нулевата хипотеза за липса на такава разлика не може да бъде отхвърлена и при наличие на висока корелация, това би означавало, индексите на съответните въпроси съхраняват не само относителните, но и абсолютните си позиции на съответна скала.

Дискриминативна сила (D)

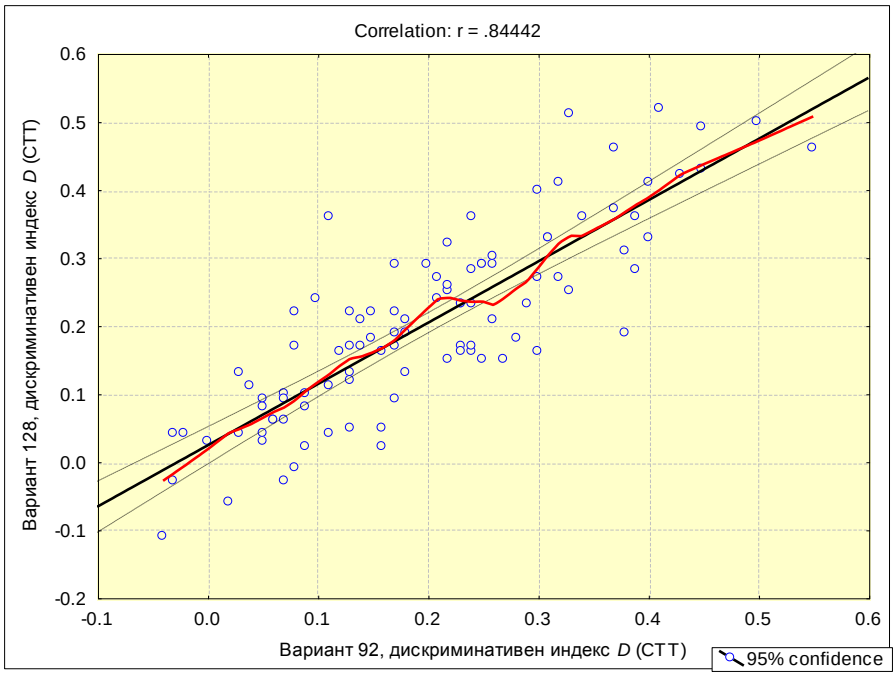
В следващата таблица са представени резултатите от корелационния анализ на индексите на дискриминативна сила на въпросите, изчислени по СТТ за съответните двойки тестови варианти. Както беше отбелязано, поради ранговия характер на индекса, като мярка за неговата стабилност е използван коефициентът на рангова корелация R на Спиърмън. Успоредно с това са изчислени и съответните коефициенти на линейна корелация. В последната колона е представена статистическата значимост на получените корелационни коефициенти.

Таблица 36. Коефициенти на стабилност (*R* и r_{xy}) на индекса на дискриминативна сила *D* в рамките на CTT

Двойка тестове	Тестов вариант	Коефициенти на стабилност (*R*, r_{xy})	Статистическа значимост (*p*)
1.	вар. 92	R = 0.827	$p < 0.05$
	вар. 128	r_{xy} = 0.844	$p < 0.05$
2.	вар. 96	R = 0.820	$p < 0.05$
	вар. 132	r_{xy} = 0.819	$p < 0.05$
3.	вар. 146	R = 0.796	$p < 0.05$
	вар. 110	r_{xy} = 0.817	$p < 0.05$

Данните в таблицата се характеризират с някои интересни особености. Коефициентите на стабилност на D (R на Спиърмън) при трите двойки тестови варианта са положителни, високи и статистически значими на ниво $p < 0.05$. Те се намират в сравнително тесния интервал от 0.80 до 0.83, който е разположен далеч над приетата прагова стойност от 0.70. Същевременно наблюдаваните стойности са близки по големина, което показва, че устойчивостта не е единично явление. Интересно е да се отбележи, че равнищата на двата типа корелационни коефициенти – на рангова и на линейна корелация, при всяка двойка варианти са твърде близки. При една от двойките тестови варианти (96 – 132) стойността на ранговия коефициент е по-висока от тази на линейния, но при останалите два от случаите (варианти 92 – 128 и варианти 146 – 110) коефициентите на рангова корелация дори са малко по-ниски тези на линейна корелация. Следователно във взаимовръзката между тези индекси се наблюдава ясно изразен линеен компонент.

Фигура 22. Разсейване на индексите на дискриминативна сила *D* на въпросите от варианти 92 и 128



Свидетелство за линейния характер на взаимовръзката между индексите на дискриминативна сила е горната диаграма на разсейването на техните стойности от двойката варианти 92 и 128.

За апроксимиране на точките са приложени два модела – линеен и нелинеен. Приложеният нелинеен регресионния модел е известен като локално претеглена регресия (*locally weighted scatterplot smoothing*, *LOWESS*). Тя се определя за всяка точка (обект) и за най-близките до него точки. Приема се, че този метод води до по-добро представяне на формата на връзката между съответните две променливи. Функцията, свързваща индексите на дискриминативна сила на въпросите от двата тестови варианта, определена по този метод, може да се определи като монотонно нарастваща. В един от сегментите (в интервала 0.22 -0.26) по хоризонталната ос се наблюдава дори обратна тенденция, към понижаване на стойностите по вертикалната ос с увеличаване на тези по хоризонталната.

Като цяло обаче на горната графика може да се забележи ясно изразения линеен характер на взаимовръзката между стойностите на D на въпросите в двата теста. Той се изразява във формата и ориентацията на облака от точки, представящи въпросите, както и от сравнително слабото им разсейване. Въпросите са групирани около регресионната линия, макар че има само някои отделни въпроси, чиито координати ги поставят извън общата маса на останалите. Така например над регресионната линия се открояват два въпроса с номера 24 и 88 с координати (стойности на дискриминативния индекс) съответно (0,11; 0,36) и (0,33; 0.51. Под линията като изключение се очертава въпрос 15 със стойности (0,38; 0,19). Отстраняването само на тези три въпроса би повишило стойността на R от 0.827 на 0.851, а на r_{xy} – от 0.844 на 0.875. Следва да обърнем внимание и на още една особеност. В два от сегментите на графиката на монотонната функция, в интервалите -0.04 – 0.17 и 0.31 – 0.55, нейната форма е изгладена и почти съвпада с линейната регресионна права.

Ще направим още една оценка на формата на връзката между двете оценки на D чрез проверка на хипотезата за нулева стойност на регресионния коефициент b:

$$H_0 : b = 0.00 .$$

Таблица 37. Оценка на регресионния коефициент за вариант 92

	b	**Стандартна грешка на b**	***B***	**Стандартна грешка на B**	***t* (97)**	***p***
свободен член			0.023	0.0138	1.699	0.092
D (вар. 92)	*0.844	0.054	*0.901	0.058	15.525	0.000

Забележка: Маркираните коефициенти са значими при p < 0.05

Нулевата хипотеза може да бъде отхвърлена на ниво α = 0.05. Данните сочат не само за високата стойност на ъгловия коефициент, но и за ниската стандартна грешка

на неговата оценка. Цялостната оценката на годността на модела, направена чрез ANOVA, е също висока – $F(1, 97) = 241.028$, $p = 0.000$. Изравненият коефициент на детерминация R^2 квадрат има стойност 0.710. Той отразява онази част от дисперсията на зависимата променлива в популацията от въпроси, която може да бъде обяснена чрез построения линеен модел. наблюдаваната стойност също е свидетелство за качеството на този модел.

Като втора мярка на стабилността на индекса на дискриминативна сила D е направена статистическа оценка на разликата между медианите на стойностите на въпросите от съответните два тестови варианта. Нулевата хипотеза е проверена чрез Знаково-ранговия T-тест на Уилкоксън за зависими извадки (*Wilcoxon matched pairs test*). На следващата таблица са представени резултатите от теста при отделните двойки варианти на ТОП.

Таблица 38. Резултати от Знаково-ранговия тест на Уилкоксън за индекса на дискриминативна сила *D*, изчислен по СТТ

Двойка тестове	Тестов вариант	Брой наблюдения	Медиана	Тестова статистика (*T*)	Статистическа значимост (*p*)
1.	вар. 92	99	0.190	1862.00	0.690
	вар. 128		0.190		
2.	вар. 96	99	0.205	1848.00	0.780
	вар. 132		0.210		
3.	вар. 146	100	0.180	1790.00	0.095
	вар. 110		0.190		

Съдейки по стойностите на тестовата статистика T и нейната статистическа значимост p, няма основания за отхвърляне на нулевата хипотеза при нито една от двойките тестови варианти. Тестът на Уилкоксън показва, че поставянето на тестовите въпроси в различни условия не предизвиква статистически значими разлики в средните равнища на техните дискриминативни индекси. Може да се приеме, че медианите на разпределенията на този индекс в съответните двойки тестови варианти са равни. Действително, дори и като извадкови статистики те имат равни (при варианти 92 и 128) или много близки стойности (при останалите двойки тестови варианти).

Бисериален коефициент на корелация (r_{bis})

В рамките на СТТ бисериалният коефициент на корелация r_{bis} се използва като втора мярка на дискриминативната сила на въпросите, наред с класическия дискриминативен индекс. За оценка на неговата стабилност са използвани същите коефициенти на рангова и на линейна корелация. Резултатите от анализа са представени в следващата таблица.

Таблица 39. Коефициенти на стабилност (R и r_{xy}) на бисериалния коефициент на корелация r_{bis} в рамките на CTT

Двойка тестове	Тестов вариант	Коефициент на стабилност (R, r_{xy})	статистическа значимост (p)
1.	вар. 92	R = 0.792 r_{xy} = 0,811	< 0.05 < 0.05
	вар. 128		
2.	вар. 96	R = 0.808 r_{xy} = 0,797	< 0.05 < 0.05
	вар. 132		
3.	вар. 146	R = 0.728 r_{xy} = 0,779	< 0.05 < 0.05
	вар. 110		

В горната таблица също се наблюдават високи рангови коефициенти на стабилност и дори по-високи от тях линейни коефициенти, които надвишават установената прагова стойност от 0.70. Сравнени със стойностите от предходния анализ на класическия дискриминативен индекс D, те се характеризират с малко по-ниски равнища. От друга страна, стойностите попадат в сравнително тесния интервал от 0.73 до 0.81, което говори за възпроизводимостта на бисериалния коефициент като мярка на дискриминативната сила на въпросите.

Вторият критерий за стабилност на бисериалния коефициент като мярка на дискриминативната сила на въпросите се основава на статистическата оценка на разликата между медианите на съответните двойки тестови варианти, верифицирана чрез Знаково-ранговия T-тест на Уилкоксън за зависими извадки. Резултатите от анализа са представени в следващата таблица.

Таблица 40. Резултати от Знаково-ранговия тест на Уилкоксън с повторни измервания за бисериалния коефициент на корелация r_{bis} в рамките на CTT

Двойка тестове	Тестов вариант	Брой наблюдения	Медиана	Тестова статистика (T)	Статистическа значимост (p)
1.	вар. 92	99	0.215	2046.500	0.304
	вар. 128		0.220		
2.	вар. 96	99	0.230	2148.000	0.624
	вар. 132		0.230		
3.	вар. 146	100	0.180	2113.500	0.537
	вар. 110		0.195		

Съдейки по тестовите статистики и асоциираните с тях равнища на статистическа значимост, нулевата хипотеза за равенство на медианите не може да бъде отхвърлена при нито една от анализираните двойки тестови варианти.

Трудност (p)

Трудността на въпросите p е характеристика, която пряко повлиява трудността на целия тест. Както беше отбелязано, в резултат на начина на неговото изчисляване в рамките на СТТ, този индекс на въпросите формира рангова скала. Поради това наблюдаваните „сурови" стойности на p бяха трансформирани в z-единици на нормираното нормално разпределение, формиращи интервална скала. Беше показано обаче, че скалата на суровите стойности на индекса p не е ординална в строгия смисъл на думата и че тя притежава „надрангови" характеристики.

Ето защо при изследването на стабилността на индекса на трудността са приложени три различни подхода за нейното определяне. Първо, скалата на трудността p е разгледана като строго ординална и за оценка на стабилността на индексите е използван коефициентът на рангова корелация R на Спиърмън. Второ, тя е третирана като интервална и за оценка на стабилността на индексите е използван коефициентът на линейна корелация r_{xy} на Пиърсън. И накрая, като основа за оценката на стабилността са използвани трансформираните в z-единици сурови стойности, върху които също е приложен коефициентът на линейна корелация. Резултатите са представени в следващата таблица.

Таблица 41. Коефициенти на стабилност (*R* и r_{xy}) на индекса на трудност (*p)* в рамките на СТТ

Двойка тестове	Тестов вариант	Коефициенти на стабилност (*R*, r_{xy})	Статистическа значимост (*p*)
1.	вар. 92	R (p) = 0.924	< 0.05
	вар. 128	r_{xy} (p) = 0,928	< 0.05
		r_{xy} z(p) = 0.930	< 0.05
2.	вар. 96	R (p) = 0.986	< 0.05
	вар. 132	r_{xy} (p) = 0,992	< 0.05
		r_{xy} z(p) = 0.990	< 0.05
3.	вар. 146	R (p) = 0.978	< 0.05
	вар. 110	r_{xy} (p) = 0,986	< 0.05
		r_{xy} z(p) = 0.985	< 0.05

Нека да разгледаме най-напред стойностите на "валидните" коефициенти на стабилност – R на Спиърмън, изчислени върху суровите стойности на p, и r_{xy}, изчислени върху стандартизираните им z-стойности. Без съмнение, и двата типа мерки свидетелстват за изключително високите равнища на стабилност на индекса на трудност на въпросите. Съответните корелационни коефициенти имат стойности над 0.90, в повечето случаи достигат 0.98 или 0.99 при нива на статистическа значимост под 0.05. Може да се отбележи, че наблюдаваните стойности на коефициента на рангова корелация са малко по-ниски от тези на линейна корелация, но разликите между тях са несъществени, едва във втория десетичен знак.

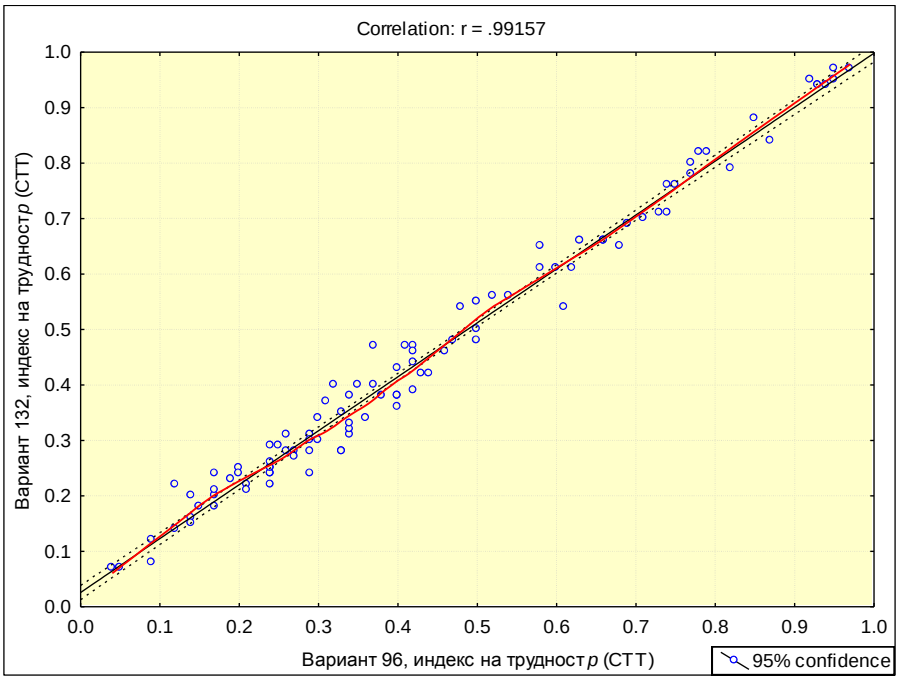
Коефициентът на стабилност, чиято валидност е под въпрос – коефициентът на

линейна корелация, приложен върху суровите стойности на p, образуващи рангова скала, достига същите равнища, както и предходните два. Нещо повече, най-високата стойност сред всички коефициенти на стабилност (0,992, при двойката варианти 96 - 132) се наблюдава именно при r_{xy}, приложен върху суровия индекс p.

Преди да разгледаме особеностите на връзката между суровите стойности на p, ще добавим, че се наблюдава известна разлика между равнищата на коефициентите на стабилност в отделните двойки тестови варианти. При първата двойка стойностите на статистиките гравитират около 0.93, а при втората и третата – около 0.98 – 0.99. Това е още едно свидетелство за взаимната заменяемост на тези мерки на стабилността.

Линейният характер на взаимовръзката на суровите (рангови) стойности на p от два тестови варианта е илюстрирана на следващата графика. Моделът е апроксимиран спрямо данните чрез стандартната линейна функция от типа $y = a + bx$, както и чрез локално претеглена регресия, използвана и за анализ на взаимовръзките между индексите D.

Фигура 23. Разсейване на суровите стойности на индекса на трудност (*p*) на въпросите от варианти 96 и 132



В горната диаграма може лесно да се забележи, че точките въпросите са разположени на (или много близо до) регресионната линия (представена с непрекъсната линия в черно). Почти липсват въпроси, чийто точки да се отклонява значително от този модел. Като вземем предвид и наклона на регресионната линия, може да заключим, че е разположена под ъгъл, който предполага много близки, почти съвпадащи стойности на индекса на трудност при двата теста. Забелязва се известно натрупване на точки в интервала 0.15 – 0.45, което съответства на големия брой въпроси с индекси на труд-

ност в този интервал.

Функцията, която описва съпоставянето на суровите стойности на трудността p от двата варианта на теста, определена по метода *LOWESS*, може да се разглежда като монотонно растяща. Нейното графично представяне е вълнообразна линия (представена с плътна непрекъсната линия в червено). Трябва да признаем обаче, че монотонната функция следва почти дословно регресионната права на линейната функция и че нейните отклонения в едната или другата посока са минимални. Забелязва се още, че тази функция е (почти) линейна в интервала 0.60 – 1.00.

Втората оценка на стабилността на индекса на трудност е направена върху стандартизираните z-стойности, с прилагане на ANOVA с повторни измервания.

Таблица 42. Резултати от дисперсионния анализ с повторни измервания за стандартизирания индекс на трудност *z*(*p*) в рамките на CTT

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (*F*)	Статист. значимост (*p*)
1.	вар. 92	0.094	1	0.094	2.769	0.099
	вар. 128					
2.	вар. 96	0.132	1	0.132	23.587	0.000
	вар. 132					
3.	вар. 146	0.000	1	0.000	0.018	0.894
	вар. 110					

Резултатите от последователните тестове на нулевата хипотеза за отделните двойки варианти не са консистентни. При първата и третата двойка нулевата хипотеза не може да бъде отхвърлена, но при втората тя може да бъде отхвърлена при равнище на p = 0.00. Следователно, средните стойности на стандартизираните индекси на вариантите 96 и 132 не са равни, макар че наблюдаваната разлика между тях (съответно 0.166 и 0.115) е само 0.052 единици.

Дали тази разлика е голяма? За да се даде отговор на този въпрос, следва да се направи оценка на размера на ефекта, което е и изискване на APA (Wilkinson, L., & APA Task Force on Statistical Inference, 1999). За оценка на размера на ефекта в случаите на повторни измервания се използва коефициентът на частна корелация ета на квадрат (*partial eta-squared*), който отразява дела на вариацията на ефекта и на грешката в зависимите променливи, която може да бъде обяснена с въздействието на съответния фактор. Взаимодействието между двата типа променливи може да се разглежда като „корелация” между тях, поради което чрез степенуването на този коефициент може да се определи „чистият” ефект.

При втората двойка размерът на ефекта е $\eta_p^2 = 0.194$, при мощност на критерия 0.998 (при α = 0.05). С други думи, 19.4% от дисперсията в стандартизираните стой-

ности на индекса на трудност може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л. Ако се придържаме към критериите на Дж. Коен за оценка на големината на ефекта, определен чрез коефициента на частна корелация η^2, следва да признаем, че ефектът на извадките от и. л. върху трудността на въпросите при тази двойка тестови варианти е твърде съществен.

2.5. Стабилност на параметрите на въпросите, определени в съответствие с Теорията за отговор на тестов въпрос

Дискриминативна сила (a)

Методите, използваните за анализ на стабилността на параметрите на тестовите въпроси, определени в рамките на IRT, се определят от интервалния характер на скалите, образувани от тях. Това са коефициентът на линейна корелация на Пиърсън и дисперсионният анализ с повторни измервания.

В следващата таблица са представени резултатите от корелационния анализ за избраните три двойки тестови варианти.

Таблица 43. Коефициенти на стабилност на параметъра на дискриминативна сила (*a*) в рамките на IRT

Двойка тестове	Тестов вариант	Коефициент на корелация (r_{xy})	Статистическа значимост (*p*)
1.	вар. 92	0.901	< 0.05
	вар. 128		
2.	вар. 96	0.888	< 0.05
	вар. 132		
3.	вар. 146	0.854	< 0.05
	вар. 110		

Съответствията между стойностите на този параметър в дойките тестови варианти са много високи, както става ясно от получените коефициенти на корелация. Те варират в границите 0.85 до 0.90 и са статистически значими на ниво $p < 0.05$. Макар че равнищата на стабилност при отделните двойки варианти се различават, всички те са над приетата прагова стойност от 0.70 и показват високата степен на съгласуваност на параметрите на дискриминативна сила.

Изненадващи обаче са резултатите от проверката на съотношенията между средните стойности на този параметър. Данните в следващата таблица дават основание нулевите хипотези да бъдат последователно отхвърлени на равнище $p < 0.05$. Това означава, че като цяло стойностите на този параметър в едната скала са изместени спрямо тези от другата със стъпка, равна на разликата между средните им стойности.

Таблица 44. Резултати от дисперсионния анализ с повторни измервания на параметъра на дискриминативна сила (*a*) в рамките на IRT

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (*F*)	Статист. значимост (*p*)
1.	вар. 92	0.111	1	0.111	88.725	0.000
	вар. 128					
2.	вар. 96	0.017	1	0.017	12.068	0.001
	вар. 132					
3.	вар. 146	0.017	1	0.017	7.443	0.007
	вар. 110					

Нека да разгледаме още една група от статистики, чрез които можем да направим допълнителна оценка на получените резултати. При първата двойка (варианти 92 – 128) разликата между наблюдаваните средни стойности е 0.047, но това води до размер на ефекта $\eta_p^2 = 0.475$, при мощност на критерия 1.000 (при α = 0.05). С други думи, 47.5 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л. При втората двойка (варианти 96 – 132) разликата между наблюдаваните средни стойности е 0.019, което води до по-малък размер на ефекта $\eta_p^2 = 0.110$, при мощност на критерия 0.931 (при α = 0.05). Поради това 11.0 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л. При третата двойка (варианти 110 – 146) разликата между наблюдаваните средни стойности е по-малка от предходната (0.018) и поради това размерът на ефекта е по-малък - $\eta_p^2 = 0.070$, при мощност на критерия 0.771 (при α = 0.05). С други думи, 7.0 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

Съгласно критериите на Дж. Коен, при първата двойка размерът на ефекта е голям, при втората – по-скоро умерено голям, а при третата - среден.

Трудност (b)

Устойчивостта на параметъра трудност на въпросите е на малко по-високо равнище, отколкото тяхната дискриминативна сила. Данните от анализа на този параметър са представени в следващата таблица.

Данните показват каква е взаимовръзката между трудността на въпросите в различните двойки тестови варианти. Тук се наблюдават най-високите коефициенти на стабилност от всички, изчислени чрез теоретичния модел на Теорията за отговор на тестов въпрос, със статистическа значимост на равнище $p < 0.05$.

Таблица 45. Коефициенти на стабилност (r_{xy}) на параметъра трудност (b) в рамките на IRT

Двойка тестове	Тестов вариант	Коефициент на корелация (r_{xy})	Статистическа значимост (p)
1.	вар. 92	0.967	< 0.05
	вар. 128		
2.	вар. 96	0.972	< 0.05
	вар. 132		
3.	вар. 146	0.915	< 0.05
	вар. 110		

От представените данни се вижда, че равнищата на устойчивост на трудността при различните двойки тестови варианти, макар и различни, се намират д интервала 0.92 – 0.97 и следователно надхвърлят праговата стойност от 0.70. В допълнение, резултатите от дисперсионния анализ, представени на следващата таблица, не дават основание за отхвърляне на нито една от нулевите хипотези за равенство на средните стойности на индекса на трудност в отделните двойки тестови варианти.

Таблица 46. Резултати от дисперсионния анализ с повторни измервания на параметъра на трудност (b) в рамките на IRT

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (F)	Статист. значимост (p)
1.	вар. 92	0.053	1	0.053	0.555	0.458
	вар. 128					
2.	вар. 96	0.114	1	0.114	1.230	0.270
	вар. 132					
3.	вар. 146	0.011	1	0.011	0.0441	0.834
	вар. 110					

Налучкване на правилния отговор (c)

Резултатите от корелационния анализ на съответствията на стойностите на третия параметър на въпросите показват високи равнища на стабилност. Корелационните коефициенти при отделните двойки варират от 0.84 до 0.90 при нива на статистическа значимост p <0.05, както сочат данните от следващата таблица.

Подобно на втората оценка на стабилността на дискриминативната сила, и тук резултатите от проверката на съотношенията между средните стойности на този параметър са в противоречие с очакваните.

Таблица 47. Коефициенти на стабилност (r_{xy}) на параметъра на налучкване на правилния отговор (*c*) в рамките на IRT

Двойка тестове	Тестов вариант	Коефициент на корелация (r_{xy})	Статистическа значимост (*p*)
1.	вар. 92	0,895	< 0.05
	вар. 128		
2.	вар. 96	0,874	< 0.05
	вар. 132		
3.	вар. 146	0,837	< 0.05
	вар. 110		

Данните в следващата таблица дават основание за последователно отхвърляне на нулевите хипотези за всяка двойка варианти на равнище $p < 0.05$. Това означава, че като цяло стойностите на параметъра за налучкване на правилния отговор при едната скала са изместени спрямо тези от другата със стъпка, равна на разликата между средните им стойности.

Таблица 48. Резултати от дисперсионния анализ с повторни измервания на параметъра на налучкване на правилния отговор (*c*) в рамките на IRT

Двойка тестове	Тестов вариант	Сума от квадратите	Степени на свобода	Среден квадрат	Тестова статистика (*F*)	Статист. значимост (*p*)
1.	вар. 92	0.000	1	0.000	4.717	0.032
	вар. 128					
2.	вар. 96	0.001	1	0.001	11.860	0.001
	вар. 132					
3.	вар. 146	0.001	1	0.001	9.558	0.003
	вар. 110					

Допълнителна оценка на получените резултати ще бъде направена чрез извеждането на разликите между средните стойности на този параметър, на размера на ефекта и на статистическата мощност на критерия при всяка двойка тестови варианти.

При първата двойка (варианти 92 – 128) разликата между наблюдаваните средни стойности е 0.002 и тази разлика води до размер на ефекта $\eta_p^2 = 0.046$, при мощност на критерия 0.576 (при α = 0.05). С други думи, 4.6 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

При втората двойка ((варианти 96 – 132) разликата между наблюдаваните средни стойности е малко по-висока (0.004), което води до по-голям размер на ефекта $\eta_p^2 = 0.108$, при мощност на критерия 0.926 (при α = 0.05). Поради това 10.8 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

При третата двойка (варианти 110 – 146) разликата между наблюдаваните средни стойности е 0.018, а размерът на ефекта е $\eta_p^2 = 0.088$, при мощност на критерия 0.865 (при α = 0.05). С други думи, 8.8 % от дисперсията в стойностите на този параметър може да бъде обяснена с обстоятелството, че те са определени въз основа на различни извадки от и. л.

Ако бъдат приложени критериите на Дж. Коен, при първата двойка размерът на ефекта клони към среден/ умерен, при втората – по-скоро голям, а при третата – малко над среден.

Изследване 4. Анализ на взаимовръзките между разноименните индекси/ параметри в рамките на един и същи теоретичен модел

Ще започнем анализа на допусканията за наличие на връзки между индексите, определени в рамките на CTT, и за отсъствие на такива връзки между параметрите, определени в рамките на IRT, чрез прилагане на подходящ за типа на съответната скала коефициент на корелация. При оценката на взаимовръзките между индексите D, p и r_{bis}, които участват в изследването със суровите си стойности, е приложен коефициентът на рангова корелация R на Спиърмън. При параметрите a, b и c е използван коефициентът на линейна корелация на Пиърсън. Резултатите, представени в следващата таблица, представляват съответните корелационни коефициенти, определени за всяка двойка статистики в рамките на отделните тестови варианти.

Таблица 49. Взаимовръзки между разноименните индекси и параметри

Тестови варианти	Индекси по *CTT* (*R*)			Параметри по *IRT* (r_{xy})		
	D - p	***D - r_{bis}***	***p - r_{bis}***	***a - b***	***a - c***	***b -c***
вар. 92	*0.324	*0.825	*0.405	*0.262	*-0.818	*-0.306
вар. 128	*0.348	*0.815	*0.398	*0.302	*-0.781	*-0.273
вар. 96	*0.402	*0.693	*0.637	*0.302	*-0.779	*-0.258
вар. 132	*0.352	*0.739	*0.531	*0.231	*-0.713	*-0.240
вар. 146	*0.512	*0.724	*0.554	*0.342	*-0.809	*-0.360
вар. 110	*0.522	*0.808	*0.535	*0.445	*-0.787	*-0.311

Забележка: Стойностите, маркирани със знака (*), са значими при p < 0.05

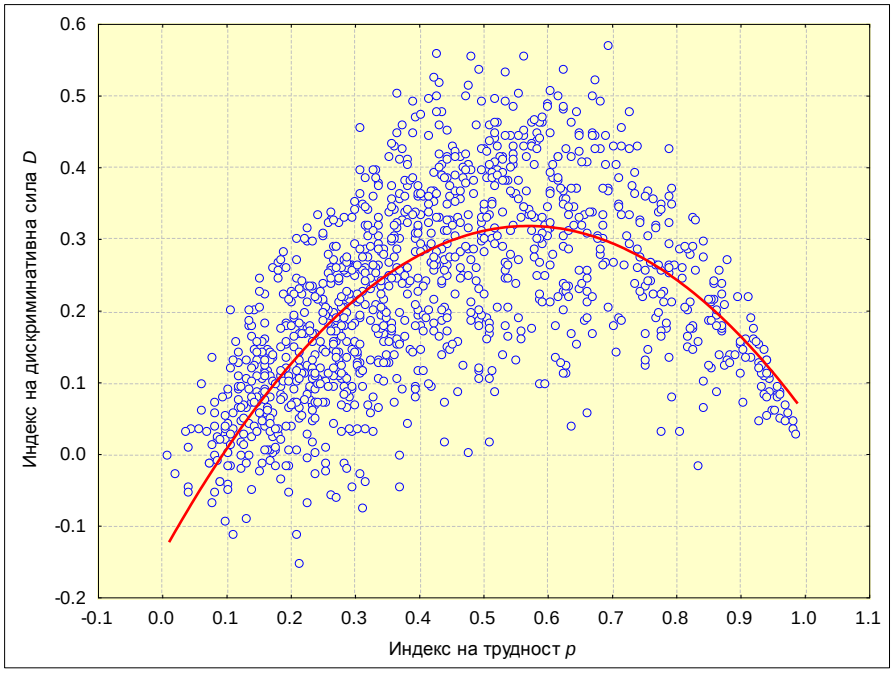
Общото впечатление от данните в горната таблицата, ако приложим класификацията на Дж. Хемфил, е, че равнището на взаимовръзките между отделните индекси е по-скоро високо. Над 85% от корелационните коефициенти имат стойности над 0.30, а някои от тях достигат до 0.83. Всички стойности са значими на ниво p < 0.05. Следва да обърнем особено внимание на значимите, високи равнища на взаимовръзка между параметрите на въпросите, определени в рамките на IRT.

2.6. Взаимовръзки между индексите на въпросите в рамките на Класическата тестова теория

Но преди това нека да разгледаме взаимовръзките между индексите на въпросите в рамките на СТТ. За отношението между индексите на дискриминативна сила D и трудност p беше изказано предположението за наличието на нелинейна връзка. За да се постави изследването на по-широка емпирична основа, беше направен повторен анализ на предполагаемата връзка чрез наблюдения върху трудността и дискриминативната сила на 1 200 въпроса, принадлежащи на 12 тестови варианта с номера 92, 96, 110, 127, 134, 141, 154, 166, 171, 175, 192 и 198. Стойностите на индексите са изчислени върху допускането, че въпросите от всеки вариант принадлежат към една скала (k = 100). На следващата графика е представена двумерна диаграмата на разсейването на въпросите.

Разсейването на въпросите дава ясна представа за това, че между двете характеристики на въпросите има връзка и че нейната форма е нелинейна по своя характер. В зоната на екстремно висока трудност (p < 0.10), както и в тази с екстремно ниска трудност (p > 0.90), въпросите се характеризират с ниски абсолютни стойности на дискриминативния индекс. В съответствие с очакванията, въпросите в средната част на скалата на трудността, особено в интервала 0.40 – 0.70, се характеризират с високи стойности на дискриминативния индекс.

Фигура 24. Взаимовръзка между трудността *p* и дискриминативната сила *D*, определени по СТТ



Прави впечатление, че отрицателни стойности на дискриминативната сила се наблюдават изключително в лявата част на хоризонталната ос (0.00 ≤ p ≤ 0.50), където

са локализирани въпросите с по-висока трудност. Негативни стойности на D в този интервал имат 51 въпроса, които съставляват 4.25% от всички наблюдения. Надясно от средата на тази ос въпроси с отрицателен индекс D (с едно изключение) липсват. Забелязва се струпване на по-голяма маса от въпроси в лявата част на хоризонталната ос, като въпросите с трудност 0.50 се намират на 64.75-тия процентил, а тези с трудност 0.60 – на 76.08-мия проценти. Подобно струпване има и в дясната част на скалата, в интервала 0.90 – 1.00, в който попадат 3.75% от въпросите.

За моделиране на взаимовръзката между променливите, формирани от двата индекса, бе приложен метода на нелинейната регресия. В ролята на зависима променлива е дискриминативната сила (D), а на независима променлива (регресор) – трудността на въпросите (p). Трябва да отбележим, че зависимостта между двата индекса е, от една страна, безспорно нелинейна, тъй като промяната (нарастването) на p е свързано с непропорционална промяна в D. От друга страна, тя не следва да се разглежда като функционална, а като корелационна, тъй като в регресионния модел не участват всички фактори, въздействат на D, което води и до несъответствия между наблюдаваните и предсказаните стойности на този индекс.

За моделиране на връзката между индексите p и D бяха проверени няколко функции (претеглен метод на най - малките квадрати, LOWESS и др.), от които бе избрана полиномна функция, зададена от следното регресионно уравнение:

$$D = (-0.143) + (1.612)\,p + (-1.411)\,p^2 \qquad (65)$$

Това е регресионен полином от втора степен, а оценката на параметрите е извършена по метода на най-малките квадрати. Изборът на тази функция за описване на връзката между двата индекса бе направен след проверка на полиномни функции с n = 2, 3, 4 и 5, а като критерий за избора бяха използвани статистическата значимости на оценките на параметрите и стандартните грешки на оценките $S_{y/x}$, които носят информация за големината на отклоненията на предсказаните от действителните стойности. При полиномните функции от трета и четвърта степен съответно един и два параметъра са статистически незначими (при този от трета степен $p(a_2)$ = 0.323, а при четвърта $p(a_0)$ = 0.955 и $p(a_1)$ = 0.824), а при полинома от пета степен нито един от параметрите не е значим при α = 0.05. При квадратичния полином се наблюдават най-ниски стандартни грешки на оценката, статистически значими стойности на оценките на всички параметри и най-тесни доверителни интервали, както е видно от следващата таблица.

Описаната параболична форма на зависимост между двата основни индекса на трудност и дискриминативна сила позволява да се предсказват стойностите на D по наблюдаваните стойности на p. Основание за това дава и високата стойност на кое-

фициента на корелация R, който отразява степента на свързаност между двата индекса. Коефициентът R се определя като втори корен от R^2 и в това изследване той има стойност R = 0.686.

Таблица 50. Оценки на параметрите на квадратичния полином

Параметри	Оценка	Стандартна грешка	*t* (119)	p	Долна граница на довер. инт.	Горна граница на довер. инт.
a_0	-0.143	0.011	-12.529	0.000	-0.166	-0.121
a_1	1.612	0.052	30.844	0.000	1.509	1.714
a_2	-1.411	0.051	-27.461	0.000	-1.512	-1.310

Забележка: Посочените граници са на 95% доверителен интервал при α = 0.05

Интересна е градацията на силата на различни типове корелационни коефициенти за съвкупността от 1 200 тестови въпроса, представени на горната графика. Докато Пиърсъновият коефициент на линейна корелация, е r_{xy} = 0.369, то стойността на коефициента на рангова корелация R е 0.478, а на корелационното отношение η за същите данни е 0.686. Това означава, че нелинейните модели са по-подходящи за описание на данните от линейния. Квадратът на нелинейния коефициент η^2 (ета на квадрат) е мярка за това каква част от дисперсията на зависимата променлива (условно това е дискриминативният индекс D) може да бъде обяснена чрез независимата променлива (условно индексът на трудност p). При анализираните данни η^2 = 0.471. Изразен в проценти, този индекс означава, че 47.10% от дисперсията на индекса D може да бъде обяснена чрез индекса p.

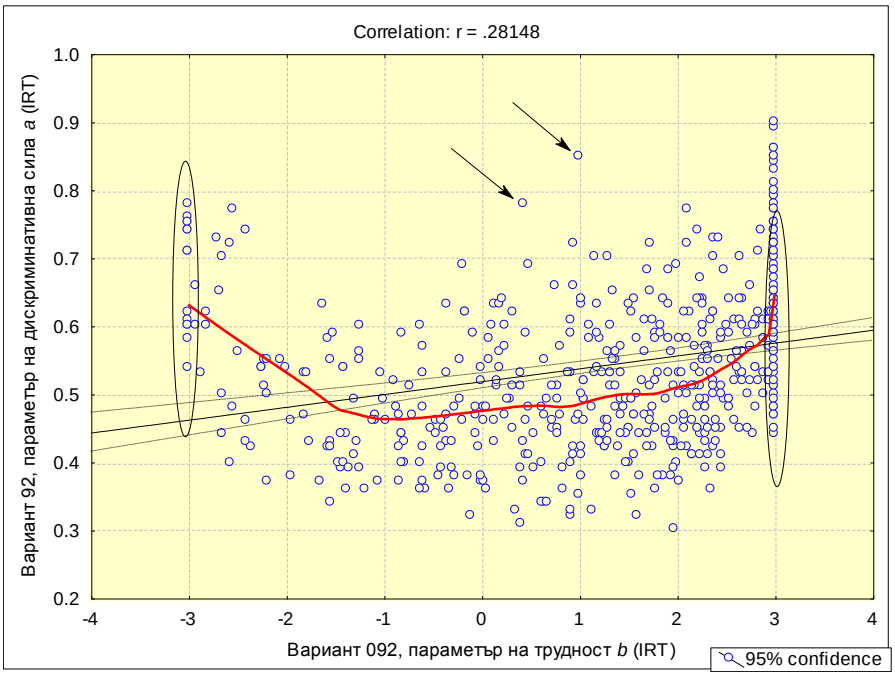
В случая, η^2 е по-скоро мярка за това доколко стойностите на p могат да се използват за прогнозиране на стойностите на D. Тъй като η^2 за нелинейните отношения има интерпретация, аналогична на тази на r^2 за линейните, разликата η^2 - r^2 би показала каква е степента на нелинейност в съвместното поведение на двете променливи (Калинов, 2010). По данните от вариант 92 степента на нелинейност η^2 - r^2 = 0.471 – 0.136 = 0.335.

2.7. Взаимовръзки между параметрите на въпросите в рамките на Теорията за отговор на тестов въпрос

Нека да се обърнем към взаимовръзките между параметрите на въпросите, определени по IRT. Особен интерес представлява връзката между дискриминативната сила a трудността b, която може да бъде разгледана от една двойна перспектива. От една страна, това е очакването за липса на връзка между двата параметъра, а от друга – установената криволинейна връзка между съответните статистики в CTT. Корелационните коефициенти в таблица 49 говорят за наличието на умерено висока до силна (по скалата на Дж. Хемфил) взаимовръзка между параметрите a и b, като стойностите

при отделните тестови варианти варират с долна граница r = 0.23 (вариант 132) достигайки до r = 0.45 (вариант 110), всички значими при α = 0.05. Наблюдава се, следователно, ясно изразена позитивна, линейна взаимовръзка. Но дали характерът ѝ е чисто линеен, както бихме могли да предположим, водейки се от интервалния тип на скалата на двата параметъра? Анализът на диаграмите на разсейване на стойностите на двата параметъра поднасят поредната изненада. При всички анализирани тестови варианти се наблюдава, повече или по-малко, ясно изразен криволинеен модел на тази взаимовръзка. Ще илюстрираме нейния характер с диаграмата на разсейването, в която са агрегирани данните за a и b на шестте анализирани варианта, с общ брой от 598 валидни въпроса.

Фигура 25. Диаграма на разсейване на стойностите на *a* и *b*



На графиката се забелязва струпване на по-трудни въпроси в дясната част на графиката (b > 0.00), характерно и за оценяването на тази характеристика чрез методите на СТТ. Интересно е „сплескването” на облака от точки отдясно, което се изразява в подреждането на серия от въпроси в дясната част на графиката с трудност b = 3.00 (маркирани в овал). Подобен ефект се наблюдава и в лявата част на графиката, в която има по-малък брой въпроси с трудност b = -3.00 (също маркирани в овал). Това се дължи на особеност в алгоритъма на психометричния софтуер, който ограничава трудността до посочените гранични стойности. Забелязват се и две нетипични стойности (*outliers*), които имат такива координати, че отстраняването им би повишило слабо линейната корелация до от 0.231 до 0.248.

По-важна особеност на тази диаграма е контурът на облака от точки, който свидетелства за нелинейния характер на връзката между двата параметъра. Първона-

чално данните са апроксимирани по метода LOWESS, а получената крива линия подсилва впечатлението от визуалния анализ. По-нататък върху данните бе приложен полиномен модел с n = 2, 3, 4 и 5, а като критерий за избора на подходяща функция бяха използвани статистическата значимости на оценките на параметрите и стандартните грешки на оценките $S_{y/x}$.

При апроксимирането с различни полиномни функции от втора, трета и пета степен бяха наблюдавани параметри, за които може да се предполага, че имат нулева стойност (p > 0.05). Единствено при функция от четвърта степен всички параметри са значими на ниво α = 0.05, с ниски стандартни грешки и тесни доверителни интервали. Данните от анализ са представени на следващата таблица.

Таблица 51. Оценки на параметрите на полиномната функция от 4-та степен

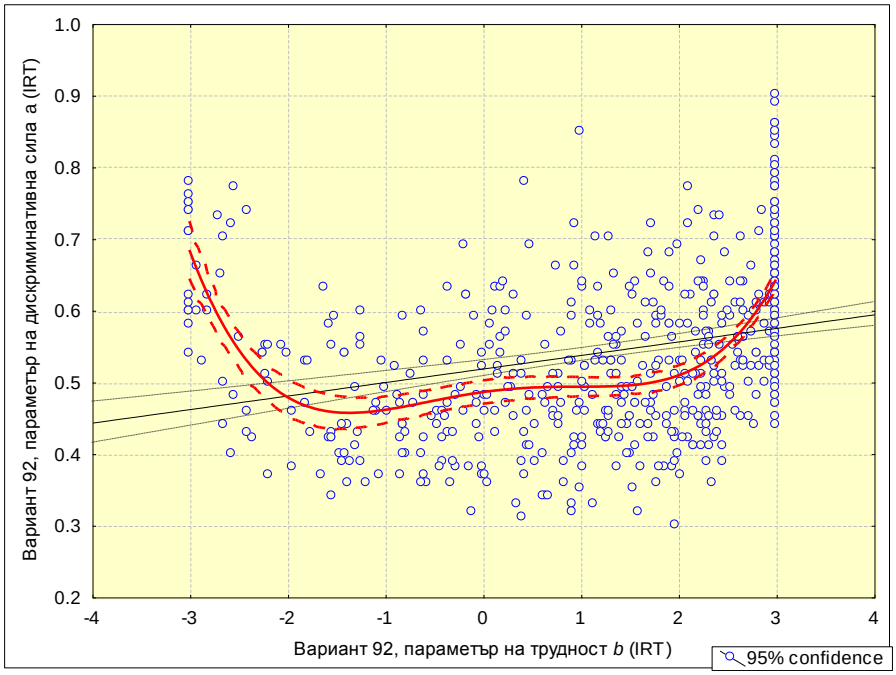
Параметри	Оценка	Стандартна грешка	*t* (119)	*p*	Долна граница на довер. инт.	Горна граница на довер. инт.
a_0	0.485	0.008	58.554	0.000	0.469	0.502
a_1	0.019	0.007	2.865	0.004	0.006	0.032
a_2	-0.012	0.005	-2.325	0.020	-0.022	-0.002
a_3	-0.003	0.001	-3.016	0.003	-0.005	-0.001
a_4	0.004	0.001	6.235	0.000	0.002	0.005

Забележка: Посочените граници са на 95% доверителен интервал при α = 0.05

Полиномната функция се задава от следното регресионно уравнение:

$$a = (0.485) + (0.019)b + (-0.012)b^2 + (-0.003)b^3 + (0.004)b^4 \quad (66)$$

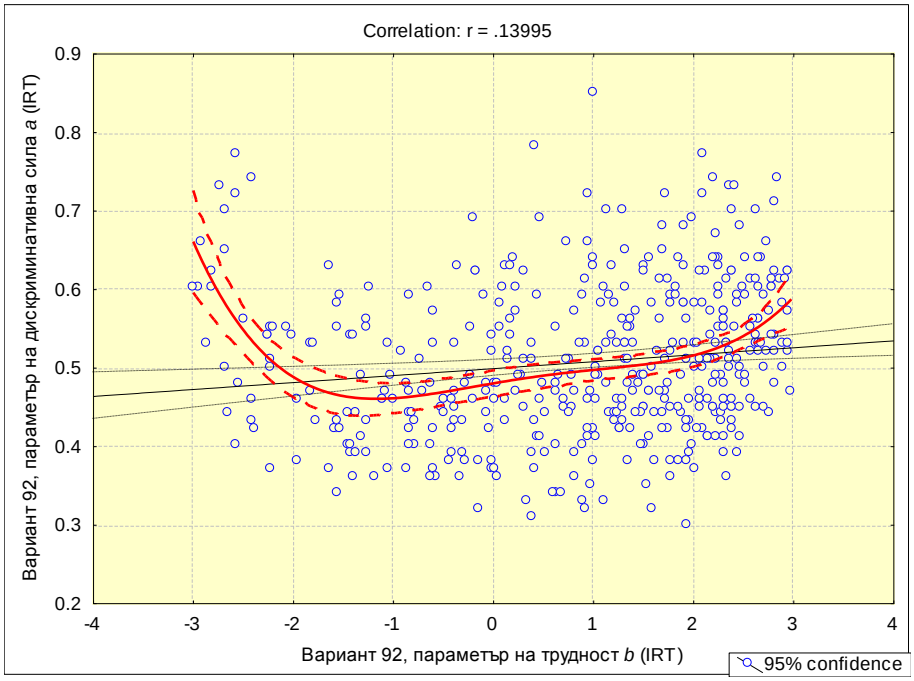
Фигура 26. Взаимовръзка между трудността (*b*) и дискриминативната сила (*a*)



На горната фигура е представена графиката на криволинейната полиномна функция, придружена от регресионната права на линейния модел на взаимовръзката между двата параметъра по данни от вариант 92. Дадени са и 95% доверителни интервали около двете линии. Ясно се забелязва тенденцията за рязко намаляване на стойностите на дискриминативния параметър (a) в интервала (-∞; -2.00), след което следва по-изгладена част в интервала (-2.00; +2.00), с тенденция към повишаване на стойностите на a и локален максимум в точка b = 0.00, следвана рязко повишаване на стойностите на дискриминативния параметър (a) дясната част на скалата, в интервала (2.00; +∞). Следователно, нелинейният характер на зависимостта между двата параметъра се проявява най-вече в двата края на континуума, в зоните на високите (отрицателни или положителни) стойности на параметъра b. Частта от криволинейната графика в интервала (-2.00; +2.00), в която тя има по-добре изразен линеен характер, следва наклона на регресионната права, съответстващ на коефициент на линейна корелация r_{xy} = 0.281, изчислен върху данни от анализираните 598 валидни въпроса. Необходимо е да се отбележи, че този модел е в голяма степен устойчив и, с известни изменения, се наблюдава при всички изследвани тестови варианти.

По-горе обърнахме внимание върху наличието на множество въпроси с трудност b = -3.00 и b = 3.00. С такава екстремно ниска стойност са 13 въпроса (2.17% от анализираните 598 въпроса), а с екстремно висока стойност са 135 въпроса (22.50% от всички). Основателно е да предположим, че „изкуственото" ограничаване на стойностите на b в интервала (-3.00; 3.00) води до изопачаване на силата и на действителната форма на взаимовръзката между двете променливи.

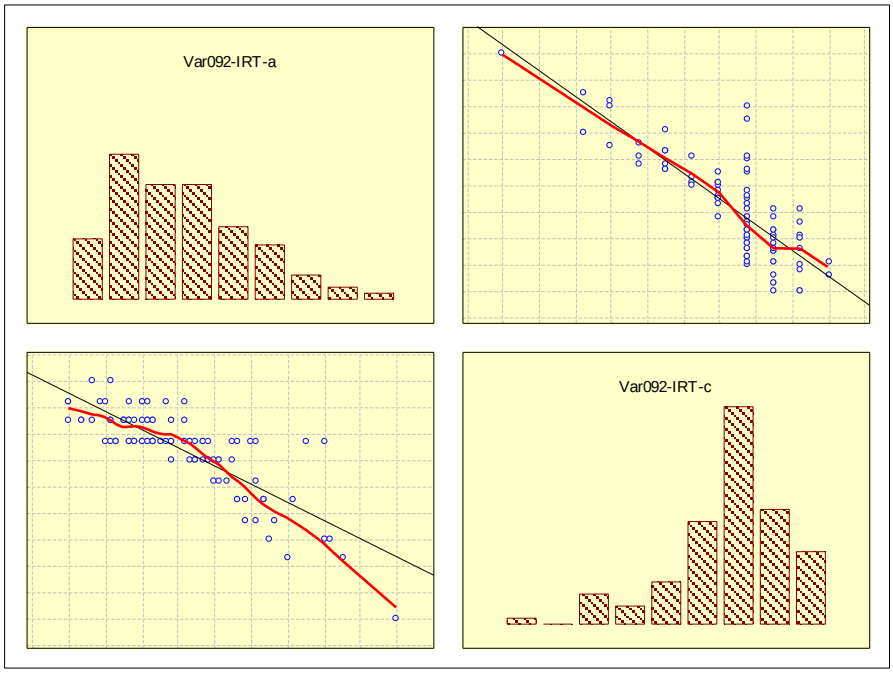
Фигура 27. Взаимовръзка между трудността (*b*) и дискриминативната сила (*a*), определени по IRT, след изваждане на въпросите с *b* = ± 3.00



Действително, след изключване от анализа на екстремните стойности на параметъра трудност (b), равни на -3.00 и на 3.00 (N=450), стойността на корелационния коефициент пада от r_{xy} = 0.281 на r_{xy} = 0.140 при $p < 0.05$. Конфигурацията на точките обаче продължава да говори за нелинейна връзка между двата параметъра, както се вижда от следващата графика, на която данните са апроксимирани с полиномна функция от четвърта степен.

Не по-малко интересни са получените данни за силни корелационни взаимовръзки между параметрите на дискриминативна сила a и налучкване на правилния отговор c. Корелационните коефициенти при различните тестови варианти варират от -0.713 при вариант 132 до -0.818 при вариант 92, при ниво на значимост α = 0.05. Получените резултати говорят на негативна корелация между тези два параметъра – с увеличаване на вероятността от налучкване на правилния отговор дискриминативната сила намалява. Взаимовръзката между тези параметри може да се разглежда като линейна, макар че при всички диаграми на разсейване се наблюдава слабо изразена нелинейност, както е показано на следващата графика (долу вляво).

Фигура 28. Диаграми на разсейването и хистограми на параметрите на дискриминативна сила (*a*) и налучкване на правилния отговор (*c*) върху данни от вариант 92

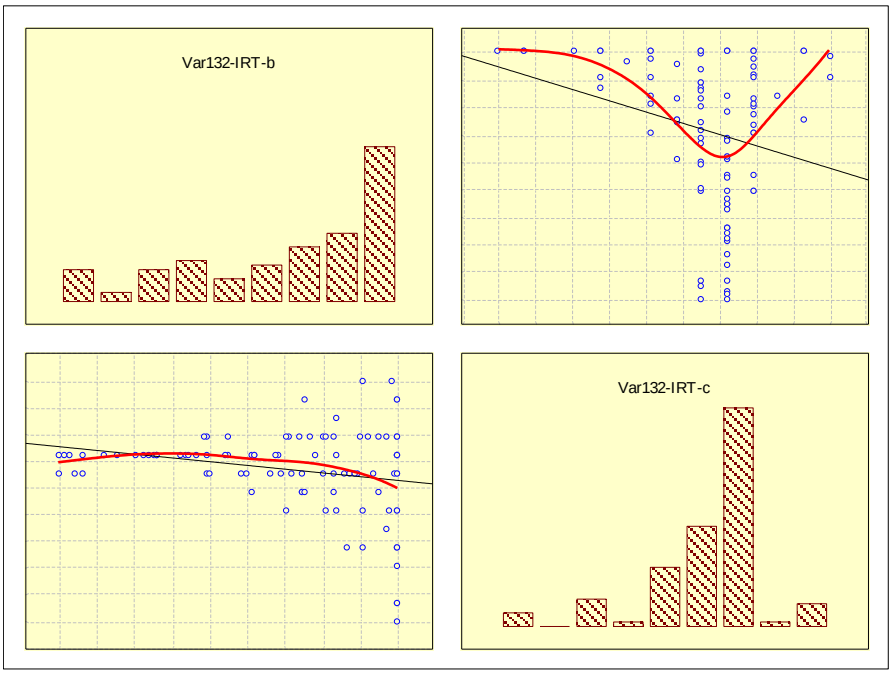


Интересно е, че тези особеност се наблюдава винаги и само тогава, когато условният „предиктор” е дискриминативната сила a, а „зависима променлива” е параметърът за налучкване c. При смяната на местата на тези променливи на двете оси на диаграмата (горе вдясно) характерът на взаимовръзката е по-скоро линеен.

Наблюдават се и сравнително високи равнища на взаимовръзка между параметрите на трудност b и налучкване на правилния отговор c. Макар и по-слабо, откол-

което с дискриминативния параметър, параметърът на налучкване c корелира с трудността на въпросите на равнища от -0.240 при вариант 132 до -0.360 при вариант 146, при ниво на значимост $\alpha = 0.05$. За взаимовръзката между тези два параметъра също е характерна обратнопропорционалната зависимост - с увеличаване на вероятността от налучкване на правилния отговор трудността намалява. Тук може да се прокара още един паралел с резултатите от предходния анализ. Диаграмите на разсейване на точките не са симетрични при смяна на местата на двата параметъра като условни „предиктори" и „зависими променливи". Променя се контурът на множеството от точки, а заедно с това и функцията, която може да се използва за апроксимация.

Фигура 29. Диаграми на разсейването и хистограми на параметрите на трудност (*b*) и налучкване на правилния отговор (*c*) върху данни от вариант 92



Изследване 5. Анализ на съгласуваността между статистиките на въпросите, определени в рамките на CTT, и кореспондиращите им статистики в рамките на IRT

2.8. Съгласуваност между статистиките на трудността

Може би най-важният аспект от съпоставителния анализ на очакваните характеристики на двете психометрични теории е проучването на степента на съгласуваност между функционално сходните им статистики. Такива са мерките за дискриминативност/ наклон и трудност/ позиция на тестовите въпроси. Като метод за изследване е използван корелационният анализ, но очевидно и тук не става дума за разглеждане на взаимовръзки от корелационен тип между едноименните статистики, а за оценка на това дали въпросите запазват своите позиции спрямо останалите в качеството им на

индекси или параметри.

Ще започнем с общ преглед на коефициентите на линейна корелация, представени на следващата таблица.

Таблица 52. Коефициенти на линейна корелация между двойки едноименни статистики

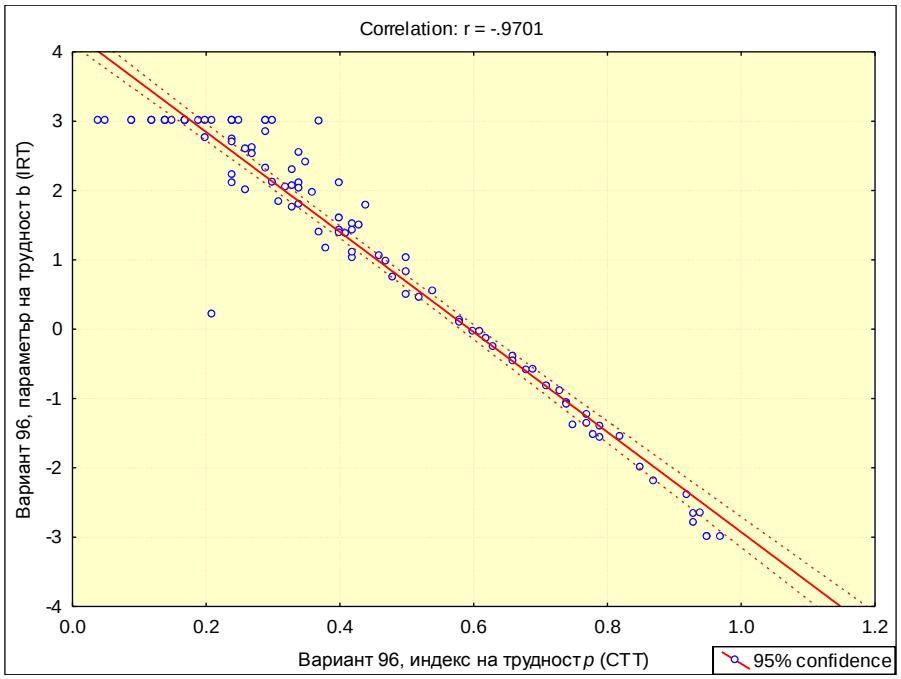
Вариант на теста	*p - b*	*D - a*	r_{bis} - *a*
вар. 92	*-0.945	-0.161	-0.064
вар. 128	*-0.970	-0.019	0.167
вар. 96	*-0.970	*-0.294	0.012
вар. 132	*-0.982	-0.099	*0.253
вар. 146	*-0.985	*-0.424	0.011
вар. 110	*-0.909	*-0.289	0.020

Забележка: Стойностите, маркирани със знака (*), са значими при p < 0.05

На първо място прави впечатление контрастът между равнищата на стойностите, описващи съгласуваността между статистиките на трудността (p - b) и тези, описващи дискриминативната сила на въпросите. Корелационните коефициенти на съотношенията между p и b в различните тестови варианти са на равнища над -0.90, достигат до -0.98, като всички стойности са значими при ниво на α = 0.05. Очакван е отрицателният знак пред тях, както имаме предвид начина на изчисляване на трудността в рамките на СТТ. Обратно, резултатите от съпоставянето на оценките на дискриминативната сила (D – a и r_{bis} - a) са неконсистентни и противоречиви. Корелационните коефициенти са сравнително ниски, при това малка част от тях са статистически значими, а при по-голямата част нулевата хипотеза не може да бъде отхвърлена.

Не по-малко интересни са формите на взаимовръзка между променливите. При анализа на диаграмите на разсейване на статистиките на трудността (p - b) от различните субтестове беше установено, че формата на взаимовръзка може еднозначно да се определи като линейна, както е показано на следващата графика върху данни от вариант 96. На графиката се забелязва струпването на точки в интервала 0.20 - 0.40 по хоризонталната ос p, съответно в интервала 1.00 – 3.00 по вертикалната ос b, което свидетелства за преобладаващата трудност на въпросите в този вариант. Може да се забележи и отклонението от регресионната линия на няколко въпроса в интервала 0.90 – 1.00 по хоризонталната ос p, каквото се забелязва и при останалите тестови варианти. Като цяло обаче линейният характер на отношенията между двете статистики не буди съмнение.

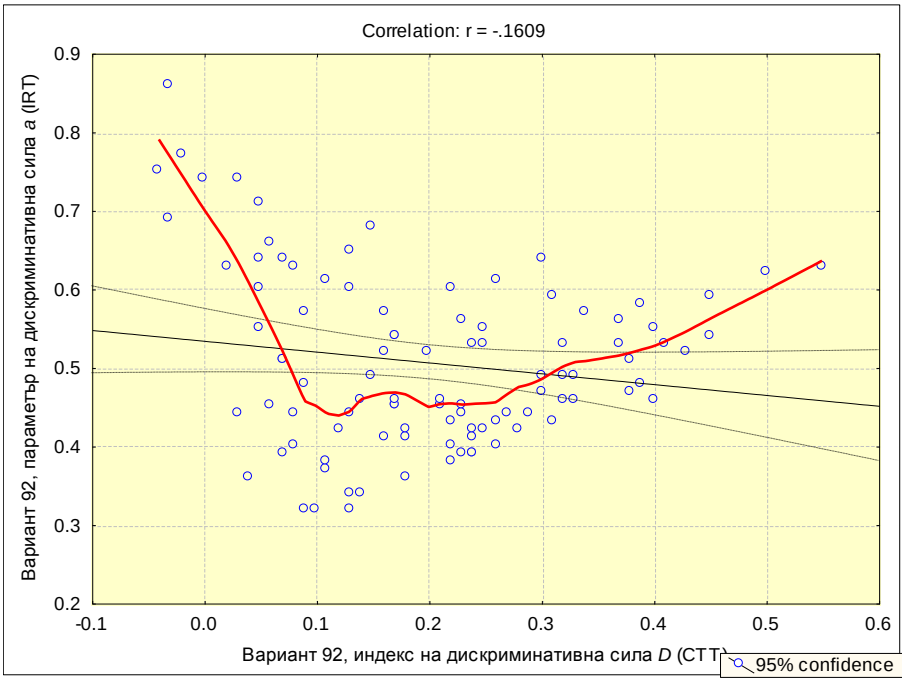
Фигура 30. Съгласуваност между статистиките на трудност *p* и *b* по данни от вариант 96



2.9. Съгласуваност между статистиките на дискриминативна сила

Не такава е картината при останалите двойки статистики на дискриминативната сила на въпросите (D – a *и* r_{bis} – a). Анализът на диаграмите на разсейването дава основание да се мисли, че сравнително ниските коефициенти на линейна корелация говорят не толкова за ниска степен на съгласуваност между техните стойности, а за нейния по-скоро нелинеен характер.

Фигура 31. Съгласуваност между статистиките на дискриминативна сила *D* и *a* по данни от вариант 92



На горната графика нелинейния тип взаимовръзка е демонстриран върху данни за D по CTT и a по IRT от вариант 92. За оценка на силата на нелинейната взаимовръзка между двете променливи беше използвано корелационното отношение ета (η). Неговата стойност по данните от вариант 92, представени на тази графика, е $\eta = 0.644$. Това означава, че между двете оценки на дискриминативната сила се наблюдава силна нелинейна корелация.

Квадратът на нелинейния коефициент η^2 (ета на квадрат) е мярка за това каква част от дисперсията на зависимата променлива (условно това е дискриминативният параметър a може да бъде обяснена чрез независимата променлива (условно дискриминативният индекс D). При анализираните данни $\eta^2 = 0.414$. Изразен в проценти, този индекс означава, че 41.40% от дисперсията на параметъра a може да бъде обяснена чрез индекса D. В случая, η^2 е по-скоро мярка за това доколко стойностите на D могат да се използват за прогнозиране на стойностите на a. Тъй като η^2 за нелинейните отношения има интерпретация, аналогична на тази на r^2 за линейните, разликата $\eta^2 - r^2$ би показала каква е степента на нелинейност в съвместното поведение на двете променливи. По данните от вариант 92 степента на нелинейност $\eta^2 - r^2 = 0.414 - 0.026 = 0.388$.

Впрочем, при диаграмите на разсейването на почти всички останали тестови варианти се наблюдава само един ясно изразен локален минимум на функцията LOWESS. Поради това графиката на съвместното вариране на статистиките D и a може да бъде разгледана като съставена от две части, във всяка от които се наблюдава, както и на горната графика, сравнително ясно изразена линейна взаимовръзка. Наляво от локалния минимум корелацията е негативна, а надясно от него – позитивна.

Таблица 53. Коефициенти на линейна корелация между статистиките *D* и *a* наляво и надясно от локалните минимуми

Вариант	Локален минимум в т. …	r_{xy} наляво от лок. минимум	r_{xy} надясно от лок. минимум
092	D=0.12	*-0.718	*0.367
128	D=0.16	*-0.512	*0.518
096	D=0.26	*-0.506	*0.471
132	D=0.21	*-0.435	*0.558
146	D=0.24	*-0.591	*0.439
110	D=0.20	*-0.678	0.260

Забележка: Стойностите, маркирани със знака (*), са значими при $p < 0.05$

Както се вижда от данните в горната таблица, минималните стойности на параметъра *a* при различните тестови варианти са в сравнително тесния диапазон от 0.12 до 0.26 от *D*. Корелационните коефициенти в последните две колони свидетелстват за наличието на висока степен на линейна съгласуваност както наляво, така и надясно от

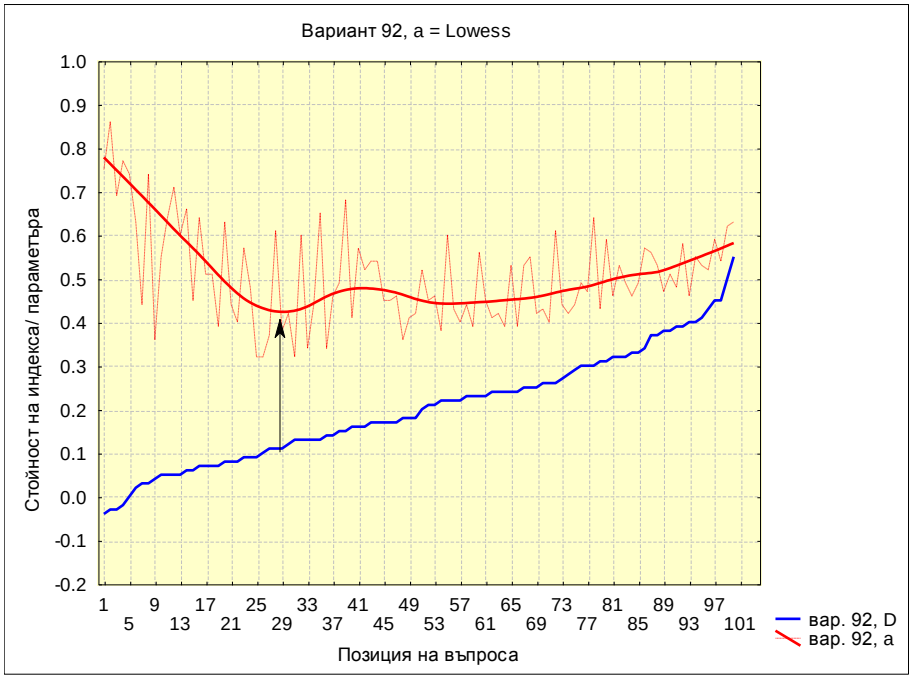
локалния минимум.

Нека да разгледаме следващата графика, на която са представени въпросите от вариант 92, подредени във възходящ ред по стойностите на индекса на дискриминативна сила *D*, и съответстващите им параметри *a*. Двете статистики са измерени в различни скали и съвместното им представяне има за цел единствено да визуализира изменението на параметъра *a* с нарастването на *D*.

Можем да забележим, че при по-голяма част от въпросите се наблюдава позитивна съгласуваност между двете статистики. Надясно от въпроса със стойност на *D* = 0.12, с нарастване на стойностите на *D* нарастват, макар и по-слабо, стойностите *a*.

При въпросите със стойности на *D*, по-ниски от 0.12, също се наблюдава съгласуваност между двете статистики, но в този интервал тя е негативна. С намаляване на стойностите на *D*, стойностите *a* нарастват, дори с по-голяма интензивност. Както показват данните в горната таблица, съгласуваността между двете статистики наляво от съответния локален минимум е по-силно изразена, с по-високи, значими коефициенти на корелация, отколкото в интервала надясно от него.

Фигура 32. Стойности на индекса *D* и параметъра *a*, сортирани по възходящ ред на *D*



3. Дискусия

Анализът на данните от тестовите варианти доведе до множество интересни, в някои случаи изненадващи резултати, които не се съгласуват с направените предположения за очакваното поведението на статистиките на въпросите. Очертаха се и ня-

кои важни тенденции, които хвърлят нова светлина върху тяхното съвместно поведение.

Първият изследователски въпрос в анализа е свързан със стабилността, инвариантността на статистиките на въпросите, оценени в рамките на двете психометрични теории. За оценка на този аспект от тяхното поведение бяха приложени два взаимно допълващи се подхода – корелационен анализ за изследване на относителната съгласуваност на стойностите на съответния индекс или параметър, получени в две различни условия, и дисперсионен анализ с повторни измервания – за оценка на съотношението между централните им тенденции при първото и второто измерване.

Най-напред следва да отбележим, че като цяло получените коефициенти на стабилност при всеки от наблюдаваните индекси и параметри, определени по двете теории, се характеризират с много високи равнища, които варират между 0.78 и 0.99. Всички емпирични стойности са статистически значими, разположени са в интервала над фиксирания праг от 0.70 и следователно могат да се разглеждат като свидетелство, че както индексите, определени в рамките на CTT, така и параметрите, определени в рамките на IRT, се отличават с висока степен на устойчивост, на стабилност по отношение на относителните им позиции при последователно оценяване в различни условия. Следва да се отбележи, че коефициентите на стабилност на всеки индекс или параметър, независимо върху коя двойка от тестови варианти са определени, се характеризират с близки, съпоставими стойности, което е още едно свидетелство за тяхната устойчивост и възпроизводимост.

Ако съпоставим равнищата на коефициентите на стабилност на различните статистики в рамките на всяка отделна теория, можем да установим няколко характерни особености.

В рамките на CTT системно по-високи са оценките на стабилността на индекса за трудност (*p*) на въпросите. При анализираните три двойки тестови варианти коефициентите надхвърлят 0.92, а при две от тях техните стойности са на равнища около 0.98. Сред тях е и най-високата наблюдавана стойност от 0.99, при варианти 96 – 132. Следователно трудността на въпросите може да се разглежда като тяхната най-устойчивата характеристика. По-малко устойчив е индексът на дискриминативна сила с коефициенти на корелация около 0.80 – 0.84, а най-малко – бисериалният коефициент с равнища на корелация около 0.78 - 0.81.

В рамките на IRT се наблюдава градация на коефициентите на стабилност, подобна на предходната. И тук със системно по-високи равнища се отличава стабилността на параметъра на трудност (*b*) на въпросите. При всички двойки тестови варианти коефициентите надхвърлят стойности от 0.92, а при две от наблюденията достигат до 0.97. Малко по-ниски са равнищата при параметрите на дискриминативна сила (*a*) и на налучкване на правилния отговор (*c*), с коефициенти на корелация около 0.85 – 0.90.

Следователно може да се заключи, че двете тестови теории, приложени върху

данните от ТОП, са в еднакво годни да осигурят висока степен на устойчивост на относителните равнища на статистиките на въпросите, оценени в различни условия.

Доста по-пъстри са резултатите, които произтичат от прилагането на втория критерий за устойчивост на статистиките на въпросите. При индексите, определени чрез алгоритмите на СТТ, тестовите статистики и асоциираните с тях равнища на статистическа значимост не дават основания за отхвърляне на нулевите хипотези за равенство на медианите или средните стойности на съответните разпределения. Следователно, поставянето на тестовите въпроси в различни условия, при различни извадки от и. л., не предизвиква статистически значими разлики в средните равнища на техните индекси. Изключение от тази иначе добре подредена картина прави индексът на трудност (*p*) при една от двойките тестови варианти, при който се наблюдава значителен размер на ефекта при почти максимална мощност на критерия.

Изненада обаче поднася поведението на параметрите, определени в рамките на IRT. При два от тях – дискриминативна сила (*a*) и налучкване на правилните отговори (*c*), проверката на нулевите хипотези доведе до последователното им отхвърляне при всички двойки тестови варианти. Последвалата проверка за размерите на съответните ефекти, направена в съответствие с граничните стойности на Дж. Коен, дава основание те да бъдат оценени като средни или големи, при високи стойности на мощността на критерия, в повечето случаи надхвърляща 0.80. При тези параметри на въпросите може да се говори за отместване на наблюдаваните стойности на едното разпределение в сравнение с другото. Единственият параметър, който демонстрира стабилност, е този на трудността (*b*). Проверката на нулевите хипотези при трите двойки тестови варианти води до последователното им отхвърляне.

Може да се направи заключението, че по отношение на втория критерий двете тестови теории, приложени върху данните от ТОП, не са равностойни. Прилагането на алгоритмите на СТТ води по-често и при повече статистики на въпросите до по-устойчиви резултати, отколкото тези на новата психометрична теория.

Ако съпоставим резултатите от анализите на едноименните статистики от двете психометрични теории, можем да обобщим, че най-устойчива статистика е трудността на въпросите. Макар че наблюдаваните коефициенти на корелация при отделните двойки варианти се различават, те се движат в един и същи интервал, непосредствено под максималната стойност от 1.00. Следователно, без значение коя тестова теория ще бъде приложена върху данните от ТОП, можем да очакваме най-ниска степен на вариативност при оценките на трудността на въпросите. Без да му придаваме поголямо от необходимото значение, ще отбележим обстоятелството, че като цяло равнищата на стабилността на индексите по СТТ са по-високи от тези на параметрите по IRT, както и това, че най-високата наблюдавана стойност от 0.992 е между две серии от индекси на трудността, определени по СТТ.

Статистиките на дискриминативната сила са по-малко устойчиви от тези на

трудността. В това отношение оценките на стабилността на параметрите по IRT са малко по-високи от тези на класическия дискриминативен индекс, който от своя страна има по-високи равнища от статистическия бисериален коефициент.

Ако разгледаме устойчивостта на статистиките на въпросите само от позицията на първия критерий, то можем да обобщим, че двете психометрични теории осигуряват в еднаква степен възпроизводимостта на относителните им позиции и в този смисъл са взаимнозаменяеми. Но ако приложим възприетия по-горе конюнктивен критерий за оценка, следва да посочим, че CTT „осигурява" стабилността на индексите на въпросите в 8 от наблюдаваните 9 случая (двойки тестови варианти), а IRT – в 3 от тези случаи. Следователно можем да направим заключението, че допускане 1 (а) за зависимост на индексите на трудност (p) и на дискриминативна сила (D, r_{bis}), определени в рамките на CTT, от извадките, въз основа на които са получени, не се потвърждава. Не намира опора в реалните данни и допускане 1 (б) за независимост на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на IRT, от извадките, въз основа на които са получени.

Но дали индексите на въпросите по CTT са действително независими от извадките и какъв е психометричният смисъл на получените резултати? Нека да се обърнем към операционалните дефиниции на класическите индекси, както и към формулите за тяхното изчисляване. Те са пряко свързани с извадките, въз основа на които са определени техните стойности и поради това получените резултати хвърлят светлина и върху разпределението на съответната способност в извадките. Индексът на трудност (p) отразява относителния дял на правилните отговори на даден въпрос в извадката и неговата стабилност следва да се интерпретира като свидетелство, че този относителен дял се съхранява и възпроизвежда в различни тестови сесии, при различни групи от и. л. Дискриминативният индекс (D) е производен на предходния и отразява относителния дял на правилните отговори в "силната" и слабата" група. Малко по-ниската стабилност на този индекс отразява сравнително по-високата вариативност на относителните дялове в двете групи, в което определена роля може да има и налучкването на правилните отговори. Дискриминативният индекс отразява и друга особеност на извадката – нейната хомогенност по отношение на измервания признак. По-високи стойности на индекса се получават при хетерогенни извадки (поради различията в „силната" и „слабата" група), докато хомогенните извадки биха довели до понижаване на неговите стойности. Възпроизвеждането на равнищата на този индекс при различни извадки може да се обясни със съхраняването на приблизително една и съща структура на извадките по отношение на този признак. С други думи, устойчивостта на индексите е обусловена от относително стабилната структура на извадките по отношение на измерваните способности.

Очакваната стабилност на параметрите на въпросите по IRT, тяхната независимост от извадката бе подложена на съмнение, макар че авторите твърдят, както беше

отбелязано по-горе, че стойностите на параметрите са характеристики на самите айтеми, но не и на групата, въз основа на която са изчислени. Нека да разгледаме този въпрос по-детайлно.

Характеристичната крива на въпросите се изгражда, подобно на трудността на въпросите по СТТ, въз основа на относителния дял на правилните отговори, но не за цялата група, а за всяко равнище на наблюдаваната способност, т. е. за всяка точка на скалата Θ. Теоретично, всяка промяна в относителния дял на правилните отговори в различни точки от скалата (например, 5 правилни отговора повече в точка Θ = -2 и същевременно 5 правилни отговора по-малко в точка Θ = +2) би могла да доведе до промяна във формата на тази крива, т. е. в стойностите на параметрите, които я описват. Такава промяна обаче няма да се отрази върху стойността на класическия индекс (*p*). Следователно един от източниците на вариативност следва да се потърси в чувствителността на модела към промени в структурата на извадките.

В тази връзка следва да отбележим, че твърдението за инвариантния характер на параметрите произтича от допускането, че всички извадки са извлечени от една и съща популация (Harris, 1993; Embretson & Reise, 2000; Baker, 2001). Теоретично, тези извадки могат да бъдат и непредставителни и да бъдат извлечени от левия или десния край на популационното разпределение или от централната му част, т.е. да не споделят общи (еднакви) статистически характеристики. Тези извадки обаче са функционално еквивалентни, защото биха довели до едни и същи оценки на параметрите. Оттук може да се направи предположението, че наблюдаваната вариативност на параметрите се дължи на обстоятелството, че различните извадки са извлечени от различни генерални съвкупности. Анализът на поведението на индексите по СТТ обаче подсказва, че такова предположение е по-скоро неоснователно. Следователно причините за наблюдаваната вариативност следва да се потърсят в особеностите на популационното разпределение. Както бе показано, анализът на неговата форма и размерност водят към заключението, че способностите не са разпределени нормално и не формират едномерно разпределение. А това са две от основните допускания, които формират фундамента на разглеждания модел на IRT. От друга страна, особеностите на реалните данните от ТОП не се съгласуват с тези допускания и поради това прилагането на този модел на IRT би било в противоречие с тях. Тук може да се добави и възможното влияние на обемите на извадките (за оценка на параметрите са необходими големи извадки), както и адекватността на логистичната характеристична крива към всеки отделен въпрос.

Следва да се отбележи, че числовите стойности на параметрите, получени в хода на анализа на тестовите данни, не представляват действителните им стойности, а са техни оценки. Тяхната вариативност е свидетелство за наличието на отклонения на наблюдаваните от действителните стойности. Тя не е основание да се подложат на съмнение теоретичните достойнства на изследвания модел на новата психометрична

теория. Вариативността им обаче е свидетелство за негативния ефект от несъответствието между теоретичния модел и реалните данни.

В заключение можем да отбележим, че „меката" и по-„непретенциозна" Класическа теория се справя по-добре с проблема със стабилността на статистиките на въпросите, отколкото далеч по-сложната в концептуално, структурно и математическо отношение Теория за отговор на тестов въпрос. Отговорът се крие в характеристиките на реалните данни, които се характеризират с определена степен на неподреденост и неорганизираност, на аморфност, което се оказва по-значимо препятствие пред втория, а не пред първия теоретичен модел.

Един съпътстващ, но важен от методологична гледна точка въпрос беше вплетен в оценката на стабилността на статистиките – въпросът за типа на скалата, която образува индексът на трудността (*p*) по СТТ, както и този на дискриминативна сила (*D*). Поради начина на изчисляване на техните стойности се приема, че тези скали са рангови. В съгласие с това схващане при анализите на стабилността на класическите индекси бяха приложени статистически методи, съответстващи на този тип измервания – коефициентът на рангова корелация *R* на Спиърмън и Знаково-ранговият тест на Уилкоксън за зависими извадки.

Същевременно в постановката на изследването бяха представени мненията на редица изследователи, различаващи се по това дали (1) параметричните техники са допустими за ранговите скали и (2) дали между ранговите и интервалните скали съществува рязка граница, т.е. дали обемите на тези понятия са (не)пресичащи се множества.

Ние сме склонни да се присъединим към мнението на тези, които допускат наличието на плавен преход между двата типа скали, съответно и допустимостта на параметричните техники към не-интервални скали. Това са Дж. Гайто, според когото в много психологически изследвания се борави със суб-интервални данни, П. Гарднър, за когото разликата между ординалната и интервалната скала не е черно-бяла, Р. Абелсън и Дж. Тюки, които изказват мнението, че недостигът на метрична информация не означава непременно наличието на рангова информация, а обикновено нещо повече от нея. Особено ценно е мнението на Дж. Гайто, който отбелязва, че разликата между двете скали (и техните допустими статистики) изобщо не е рязко очертана и че едни и същи данни могат да имат свойствата на две или повече скали.

Тук ще обърнем особено внимание на една оригинална идея на К. Кумбс, който предлага един нов скалов тип, наречен „подредена" метрична скала (*ordered metric scale*), която се позиционира между ранговата и интервалната. Това е скала, която предполага подреждане на обектите по някакво тяхно свойство (това е ранговият аспект на скалата), но също така и (частично) подреждане на дистанциите между тях (това е метричният й аспект) (Coombs, 1950, по Fagot, 1959 и Phillips, 1971; Coombs, 1964).

Основания да се присъединим към тази група от изследователи намираме в резултатите от направените изследвания на типа на скалата на трудността, представени по-горе. Направените оценки се базират на теоретичната постановка на П. Супес и Дж. Зинес, че типът на измервателната скала се определя от трансформацията, чрез която се преминава от една числова система към друга (т .е. от една скала към друга) (Суппес и Зинес, 1967). Анализът бе направен чрез прилагане на числови и на графични методи. Тук ще направим обобщение на резултатите, като ги представим в няколко групи.

(1) За целите на изследването суровите стойности на индекса на трудността (*p*) бяха трансформирани в *z*-единици на стандартното разпределение по алгоритъм, описан от Л. Айкен, А. Анастази и С. Урбина и др. (Aiken, 1988; Анастази и Урбина, 2001). Получените нови стойности образуват интервална скала. Очакването бе, че ако суровите стойности на индекса *p* образуват рангова скала, то взаимовръзката им със скалата на *z*-единиците следва да бъде някакъв тип монотонно, а не линейно преобразуване. Наблюдаваните коефициенти на линейна корелация при всички двойки *p* – *z*(p) имат изключително високи стойности, около и над -0.99, които са значими при $p < 0.05$. Вниманието върху високите, гранични стойностите на линейния коефициент на корелация произтича от това, че при нелинейно свързани данни той дава занижени оценки за силата на взаимовръзката (Калинов, 2010), каквито наблюдаваните очевидно не са.

Положителните резултати от направените проверки на значимостта на регресионния коефициент и на цялостната годност на линейния модел, както и стойността на изравнения коефициент на детерминация R^2, демонстрирани при един от субтестовете, вдъхват допълнителна увереност в годността на проверявания модел Следователно преходът от сурови към стандартизирани стойности на *p* е линеен и поради това скалата на суровите стойности трудността, съгласно допустимите преобразувания, може да се третира като интервална.

Анализът на диаграмите на разсейването на стойностите обаче показа интересна особеност. В един широк интервал от сурови стойности на *p* (0.10 – 0.90) преобразуването е почти строго линейно. В двата края на регресионната линия обаче (с екстремни стойности на трудността *p*, близки до 0.00 или до 1.00.) то променя своя характер в монотонно. Разбира се, тази особеност отразява неравенството на процентилите като единица на измерването. Ако разгледаме разпределението на която и да е сетия от сурови стойности, то разликите между тези стойности в процентилно изражение в средата на разпределението са много по-малки, отколкото в двата му края.

Поради това трябва да коригираме разбирането си за типа на скалата на индекса на трудността. Първо, тя не е интервална в строгия смисъл на понятието, дори и в посочения по-горе интервал. Впрочем, не само в областта на социалните и поведенческите науки, но и в други области, асоциирани с извършването на прецизни измервания, трудно би могло да бъде намерен пример за емпирична интервална скала по-

ради наличието на грешки в измерването. Второ, в нея се съдържа както метричен (линеен), така и неметричен (рангов) компонент. При това двата компонента са разграничени ясно в отделни сегменти по протежението на самата скала: тя е предимно метрична в широкия интервал от 0.10 до 0.90 и неметрична в двата края на скалата, извън този интервал.

Поради това на скалата на трудността следва да се гледа като на субинтервална, съдържаща двата типа информация, с превес на метричната информация. Като оценка за дела на неметричната информация беше използвана разликата между квадратите на корелационното отношение и линейния корелационен коефициент, която отразява степента на нелинейност в данните и която, за данните от разгледания по-горе вариант 92, възлиза на 0.013. Нищожният дял на рангова информация в тази скала се дължи на обстоятелството, че в емпиричните разпределения на индекса на трудност делът на въпросите с екстремни стойности ($p < 0.10$ и $p > 0.90$) е много малък - в изследваните тестови варианти той е средно около 4 – 5 % от въпросите.

(2) Успоредно с коефициентите на рангова корелация, за оценка на стабилността на класическите индекси бяха използвани и тези на линейна корелация. Общото впечатление е, че между оценките, направени по двата метода, се наблюдават незначителни разлики, като в почти всички случаи по-високи са равнищата на коефициентите на линейна корелация. Прилагането на тази мярка следователно не намалява, а подобрява оценката на силата на взаимовръзката между съответните две променливи.

(3) Паралелно с числените методи, за определяне на вида на корелационните връзки, съответно на типа на скалите на индексите на трудност (p) и дискриминативна сила (D), бяха приложени и графични методи. Бяха изследвани диаграмите на разсейване на суровите стойности на двата индекса, получени въз основа на данните от вариантите в отделните двойки. Анализът на диаграмите на разсейването също води към извода за интервалния характер на класическите индекси.

Към суровите стойности на класическите индекси бяха приложените два модела за апроксимация на данните – линеен и нелинеен, по метода на локално претеглената регресия. При двата индекса графичните методи, приложени паралелно, водят до почти едни и същи резултати.

При графичното представяне на криволинейните функции, описващи съвместното поведение на съответната двойка променливи, техните форми при двата индекса са изгладени, почти съвпадащи със съответните линейни регресионни прави. Само в отделни, тесни сегменти от тях взаимовръзката придобива монотонен характер, поярко изразен при дискриминативния индекс. Тъй като типът на скалата на индекса на трудност (p) беше изследван чрез отделна процедура за трансформация на суровите стойности, предположението за линейност/ интервалност на дискриминативния индекс (D) беше проверено допълнително чрез тестове за значимостта на регресионния кое-

фициент, както и за цялостната годност на модела. Резултатите потвърждават предположението за линейния (интервален) характер на скалата на дискриминативния индекс. Впрочем, и тази скала следва да се разглежда като суб-интервална, съдържаща нелинейни сегменти, но приближаваща се до интервалния тип скали.

В заключение можем да обобщим, че скалите на индексите на трудност (p) и дискриминативна сила (D), определени в рамките на СТТ, не са интервални, но не са и рангови в строгия Стивънсов смисъл на тези понятия. В неговата класификация типовете скали са теоретични конструкти или "концептуални" скали по определението на М. Бунге (Бунге, 1975). Реалните скали обаче не съответстват на концептуалните дефиниции. Обсъжданите две скали очевидно се намират в „сивата" зона между ранговия и интервалния тип скали. При това тази зона може да бъде осветлена, т.е. да се направи конкретна оценка към кой скалов тип се приближава дадена скала. Би било по-обосновано тези теоретични конструкти (техните обеми) да се разглеждат не като обичайни множества, а като размити множества, съгласно концепцията на Л. Заде (Zadeh, 1965), според която членството на даден обект (скала) към дадено множество (скалов тип) е континуално, а не дихотомично.

Вторият изследователски въпрос в анализа е свързан с взаимовръзките между разноименните индекси, от една страна, и параметри, от друга, определени въз основа на данните от отделни тестови варианти. Бяха направени две допускания за очакваното поведение на статистиките: (а) за наличие на нелинейна взаимовръзка между стойностите на индексите на трудност (p) и на дискриминативна сила (D), определени в рамките на СТТ върху една и съща извадка и (б) за липса на взаимовръзки от корелационен или функционален тип между стойностите на параметрите на дискриминативна сила (a), трудност (b) и налучкване на правилния отговор (c), определени в рамките на IRT.

Първото допускане бе обосновано чрез модела „Данни единичен стимул" на К. Кумбс (Coombs, 1964). Второто допускане следва от вероятностния подход за оценяване на параметрите.

За оценка на взаимовръзките между индексите D, p и r_{bis}, които участват в изследването със суровите си стойности, бе приложен коефициентът на рангова корелация R на Спиърмън. При параметрите a, b и c е използван коефициентът на линейна корелация на Пиърсън.

Резултатите от направените анализи свидетелстват, че между всички статистики, определени в рамките на дадена теория, се наблюдават ясно изразени взаимовръзки от корелационен тип. Ако се приложи класификацията на Дж. Хемфил, общото равнище на силата на взаимовръзките между отделните индекси и параметри може да бъде оценено като високо. Преобладаващата част от стойностите са над горната граница от 0.30, а някои от тях достигат до 0.83, като всички стойности са значими на ниво $p < 0.05$. Следва да обърнем особено внимание на високите коефициенти на взаимов-

ръзка между двете оценки на дискриминативната сила (D и r_{bis}) по CTT, както и между параметрите на дискриминативната сила a и налучкване c по IRT, достигащи до стойности от 0.70 – 0.80.

Особен интерес представляват взаимовръзките между индексите на въпросите в рамките на CTT и по-конкретно между индексите на дискриминативна сила D и трудност p, за които бе изказано предположението за наличието на нелинейна връзка. Оценките при отделните субтестове, направени чрез непараметричен коефициент на рангова корелация, са достатъчно високи (0.30+ – 0.50+), за да потвърдят това предположение – между двата индекса има поне някакъв тип монотонна връзка.

За да се разшири емпиричната основа на изследването, бе направен допълнителен анализ с данни за трудността и дискриминативната сила на 1 200 въпроса, принадлежащи на 12 тестови варианта. Интересна е градацията на равнищата на различни типове корелационни коефициенти, приложени върху тази съвкупност. Тази сила нараства от 0.369 при коефициента на линейна корелация на 0.478 при коефициента на рангова корелация, за да достигне до 0.686 при корелационното отношение η. Това означава, че нелинейните модели са по-подходящи за описание на данните от линейния.

Диаграмата на разсейването на техните стойности показва ясно, че двата индекса са свързани с криволинейна връзка, която има форма на изпъкнала парабола. Наблюдават се очакваните ниски абсолютни стойности на дискриминативния индекс в зоните на екстремно високи и ниски стойности на трудността, както и високи стойности на дискриминативния индекс в средната част на скалата на трудността.

Съвместното вариране на индексите беше апроксимирано по метода на нелинейната регресия с полиномна функция от втора степен. Беше определено и съответното регресионно уравнение с удовлетворителни оценки на параметрите. Степента на нелинейност в съвместното поведение на двете променливи (разликата $\eta^2 - r^2$), по данните от вариант 92, възлиза на 0.335.

Друга интересна характеристика, която се наблюдава в диаграмата на разсейването на двата индекса, е струпването на точки в лявата част на параболата, в посока към по-висока трудност и по-ниска дискриминативна сила на въпросите. Тук е разположена и основната маса от въпроси с отрицателни стойности на D. Този феномен заслужава по-внимателно обсъждане.

Трудността на въпросите отразява общия брой на правилните отговори на даден въпрос, следователно всяка точка, разположена по-наляво, отразява все по-намаляващия общ дял на правилните отговори. Дискриминативната сила отразява съотношението (разликата) между правилните отговори в двете екстремни групи, следователно всяка точка, разположена по-ниско, отразява все по-изравняващия се баланс между двете групи, който от даден момент нататък (D = 0.00) преминава в полза на „слабата" група. Например при въпрос 80 от вариант 192, който има най-ниската диск-

риминативна сила и е сред най-трудните в цялата съвкупност от 1 200 въпроса (D_{80} = -0.157, p = 0.216), делът на правилните отговори в „слабата" група p_{low} = 0.283, а на тези от "силната група е много по-малък - p_{high} = 0.126. Как може да бъде обяснен фактът, че лицата с ниски способности се справят с трудните въпроси еднакво добре, дори по-добре, отколкото тези с високи способности? Отговорът може да бъде потърсен в третия, останал встрани от изследването индекс на налучкване на правилния отговор. В диаграмата на разсейването се съдържат нагледни свидетелства за това, че изпитваните прилагат стратегия за налучкване на правилния отговор, на която по-нататък ще обърнем допълнително внимание. Данните дават основание да смятаме, че тази стратегия е по-присъща на лицата от „слабата", отколкото на тези от „силната" група.

Подобно струпване на точки има и в дясната част на параболата, в посока към по-голяма леснота на въпросите. Безспорно и тук с намаляване на трудността на въпросите се наблюдава изравняване на съотношението между дела на правилните отговори в двете екстремни групи, изразяващо се в намаляване на стойностите на индекса на дискриминативна сила. Те обаче не преминават граничната нулева стойност, т .е. тази тенденция не се обръща.

Следователно може да се направи обобщението, че допускането за наличие на нелинейна връзка между индексите на трудността p и дискриминативна сила D, определени в рамките на CTT, намира опора в емпиричните данни.

Неочаквани се оказват резултатите от изследването на взаимовръзките между параметрите на тестовите въпроси, определени в рамките на IRT. Тук от особен интерес е взаимовръзката между дискриминативната сила (a) и трудността (b), която може да бъде разгледана не само от гледна точка на очакването за липса на връзка между двата параметъра, но и от тази на демонстрираната криволинейна връзка между аналогичните статистики в CTT. Получените корелационни коефициенти говорят за наличието на умерено високи до високи равнища на позитивна, статистически значима взаимовръзка между двата параметъра. При това направените оценки са свидетелство за равнищата на линейния компонент въз взаимовръзките при отделните тестови варианти. Анализът на диаграмите на разсейване показа, че по-подходящ е нелинейният модел на взаимовръзката, който може да бъде представен чрез полиномна функция от четвърта степен. Особеност на сложната крива е това, че нейният нелинеен характер се проявява най-вече в двата края на континуума, в зоните на високите (отрицателни или положителни) стойности на параметъра b. В средната част на кривата, в зоната $\pm 2p$ кривата е почти изгладена, с локален максимум в точка p = 0.00, но с тенденция за повишаване на стойностите на a. Именно на тази част от кривата се дължи и линейният компонент във взаимовръзката.

Не по-малко интересна е наблюдаваната взаимовръзка между параметъра за налучкване на правилния отговор c и останалите два параметъра. При всички тестови варианти тя е негативна, статистически значима, с високи стойности, които при изс-

ледванията на взаимовръзката с дискриминативната сила *a* са системно по-високи, достигайки до равнища, надхвърлящи 0.80. При това, както и при взаимовръзката между *a* и *b*, получените оценки отразяват линейния компонент в съвместното поведение на тези два параметъра.

Диаграмите на разсейване на показват, че при двете комбинации от параметри (*a* – *c* и *b* – c) се наблюдава известна криволинейност на взаимовръзката, по-силно изразена при параметрите *b* и *c*, но, наблюдавана при различни тестови варианти, тя не се характеризира с определена устойчивост.

Следователно може да се направи заключението, че предположението за липса на взаимовръзка между параметрите на въпросите, определени чрез алгоритмите на IRT, не намира опора в емпиричните данни.

И така, параметърът на налучкване на правилния отговор *c* е негативно свързан с дискриминативната сила *a* и трудността *b*. С други думи, с увеличаване на стойността на този параметър се намаляват стойностите на останалите два. Нека да формулираме част от това твърдение по друг начин: с увеличаване на трудността на въпроса вероятността от случайно посочване на правилния отговор намалява.

Докато корелационните отношения са симетрични в статистически смисъл, т.е. $r(X, Y) = r(Y, X)$ (Калинов, 2010), то горните две твърдения не са симетрични, макар и да са еквивалентни. Смущаващо е не само това, че поведението на тези две статистики е съгласувано. Още по-смущаващо е това, че, съгласно резултатите, и. л. налучкват по-успешно коректните отговори на по-лесните въпроси, отколкото на по-трудните. Тази асиметричност може да бъде преодоляна, ако във взаимовръзките между параметъра на налучкване (*c*) и другите два се допусне съществуването не на симетрична (корелационна), а на еднопосочна (каузална) връзка.

Включването на параметъра на налучкване в уравнението на характеристичната крива на въпроса от А. Бирнбаум (Birnbaum, 1968) отразява една житейска реалност – хората, които се явяват на изпит, могат да посочат правилния отговор, дори и когато нямат необходимите знания и умения. Чрез този параметър в зависимата променлива - вероятността за правилен отговор, се включва приносът на този феномен.

Съгласно трипараметричния модел на IRT, налучкването се разглежда като вероятност за посочване на коректния отговор по случаен начин. За даден въпрос този параметър е константен за всички точки на континуума на способностите Θ, т.е. лица с различни нива на способности имат равен шанс да посочат коректния отговор. От друг страна, принципът на инвариантност на параметрите, според който те са независими от извадката, е валиден и за параметъра *c*. По-точно, параметрите са независими от разпределението на и. л. на скалата на способностите Θ. Параметрите са атрибути на характеристичната крива на въпроса, които определят нейната форма и мястото на неговото функциониране на континуума Θ. Или, както беше отбелязано по-горе, параметрите характеризират въпроса, а не групата, отговорила на въпроса (Baker, 2001).

Дали обаче параметърът на налучкване *c* е, характеристика, която произтича само от въпроса? Действително, някои особености на въпросите като броят на дистракторите, начинът на формулиране на основата на въпроса и алтернативните отговори и др., влияят пряко върху този параметър. Но налучкването следва да се разглежда и като поведение, характеризиращо даден индивид или група индивиди. Ако застанем на тази позиция, връзката между параметъра на налучкване и другите два параметъра става по-прозрачна и може да получи правдоподобно обяснение.

Ако при отговора на даден въпрос определена група от лица прибегне до стратегията на налучкване, по-вероятно е това да не са всички лица от извадката, а само тези с по-ниски способности, които се намират на левия край на континуума Θ. Свидетелство за това, че до тази стратегия прибягват по-скоро лицата от „слабата" група, отколкото на тези от „силната", беше намерено при анализа на взаимовръзката между индексите *p* и *D* по СТТ. Резултат от прилагането на тази стратегия би бил „натиск" върху лявата част на характеристичната крива отдолу нагоре, който би повишил долната й граница. Повишаването на вероятността от правилен отговор в лявата част на скалата, без да е съпроводено с подобно повишаване в дясната й част, предполага промяна в разпределението на и. л. на скалата Θ, която се изразява в отслабване на дискриминативната сила на въпроса. Графично тази промяна се проявява чрез намаляване на наклона на характеристичната крива в точка $\Theta = b$, трудността на въпроса. Това е механизмът, за който може да се предположи, че регулира съвместната вариация на двата параметъра.

Една от особеностите, свързани с въвеждането на третия параметър, е в промяната на начина на определяне на трудността, т. е. на позицията на въпроса на скалата Θ. Тя вече не е в точката, която съответства на вероятност от правилен отговор $P(\Theta) = 0.50$ (както при двупараметричния модел или при трипараметричния при $c = 0.00$), а в точка $P(\Theta) = (1 + c)/2$, т .е. средната стойност между стойността на параметъра *c* и неговия максимум 1.00. При въвеждане на третия параметър (или при промяна на стойността на *c* от 0.00 в по-висока стойност), позицията на въпроса би могла да се запази. Наличието на статистически значими негативни корелации между *c* и *b* обаче говори, че има тенденция въпросите да намаляват трудността си с увеличаване на вероятността от налучкване, т.е. да изместват позицията си наляво на скалата. Този феномен може да бъде обяснен с нарастването на общата маса на правилните отговори.

И така, може да се предположи, че силата, която отключва тази мрежа от взаимодействия, е налучкването на правилните отговори като съзнателно поведение на част от и. л. В англоезичната литература, освен термините"*guessing*" и "*pseudo-guessing*", се използва и терминът „*proneness to guessing*" в същото значение – склонност, предразположение към прилагане на такава стратегия, което може да характеризира единствено индивидите.

В този смисъл параметърът на налучкване може да бъде разглеждан, наред със способностите Θ, като втори параметър в IRT, който съдържа личностов компонент и който отразява определени психични процеси. Това схващане се отличава от традиционното, съгласно което в рамките на IRT се разглеждат две различни, несвързани групи от параметри. Първата група включва само един параметър, който описва характеристиките на индивида (личностовия параметър Θ) и втора група от 1 до 5 параметъра, които описват ситуацията, в която е наблюдаван отговорът на този индивид (параметри на въпросите).

Такова поведение на индивидите може да бъде разгледано в светлината на теорията за мотивацията. Без съмнение, то е обусловено от мотивацията за постижения, която включва множество потребности и мотиви за действие, насочени към постигането на високи резултати и значими цели. Класическият модел на мотивационния процес представя избора на дадено поведение като последица от очакването, че това поведение ще доведе до определен резултат, и от субективната желателност или ценност на очаквания резултат (Величков, 1989; Дилова, 2008). Това е популярният модел "очаквания – ценност", представен от Х. Хекхаузен.

В изпитни ситуации като тези, при които са събрани анализираните данни, са налице и двата компонента. Кандидатстудентските изпити са с висок залог, имат състезателен характер и изходът от тяхното провеждане е дихотомичен. Положителният изход не се свежда просто до постигане на високо постижение (като конкретна изпитна оценка), а до преминаване в друг социален статус – този на студент. Това е очакваният резултат от изпита и без съмнение той е желан и има висока стойност за кандидатите. В този смисъл прибягването до стратегията за налучкване, чрез която даден кандидат може да придобие определено предимство и да надхвърли действителния си бал, може да се разглежда като мотивирано от очакването, това поведение би подпомогнало постигането на желания резултат. Разбира се, налучкването не е единственото действие, предприето от кандидатите за постигане на желаната цел. Тук е предмет на анализ, защото се реализира в тесните времеви рамки на изпита и намира пряко отражение в изпитните резултати.

И. Айзен и М. Фишбайн разработват по-детайлна теория на подбудителната регулация (за планираното поведение), която е по-подходяща като обяснителен модел (Ajzen, 1991, по Дилова, 2008). В тази теория, освен фактора „атитюд към поведението" (очаквания – ценност), авторите въвеждат още два фактора – „субективна норма" и „възприеман контрол". Вторият фактор се отнася до субективно възприетата социална норма и готовността на индивида да съобрази своето конкретно поведение с нея.

Като цяло у нас социалната норма по-скоро толерира използването на, да кажем, непочтени методи за постигане на определено предимство. В една изпитна ситуация социалната норма се установява по-скоро от участниците в изпита (дори не само в конкретния изпит, а в изпитите като процедура за стратифициране), която също то-

лерира използването на такива средства като преписване и подсказване, дори и при изпити, които имат явно съревнователен характер. Нормата за приемливо поведение може да бъде зададена и от конкретен авторитет – лицето (институцията), която провежда изпита. Този авторитет би могъл да ограничи прилагането на налучкване, например под страх от редуциране на наблюдавания тестов бал с определен „коефициент на налучкване”. Можем да си представим идеална изпитна ситуация, в която всички и. л., по силата на такава императивна инструкция, не прибягват до тази стратегия. В тази ситуация лицата, които нямат необходимата компетентност да отговорят на даден въпрос(в преобладаващата си част – в левия край на скалата Θ), няма да посочат отговор и ще получат нула точки. Поради това потенциалът на въпросите за налучкване, дори и да са двуалтернативни, няма да бъде реализиран. Ще бъде реализиран обаче двупараметричен модел с c = 0.00 при всички въпроси.

Обикновено изпитващите не поставят такива ограничения, дори напротив, при изпълнението на много тестови програми изпитваните са окуражавани да посочат какъвто и да е отговор на въпросите, по които не се чувстват уверени. Макар че при провеждане на Теста по общообразователна подготовка кандидат-студентите не са явно поощрявани да налучкват, няма ограничение, което да ги възпира. Още повече, че налучкването е най-„невинното” сред непочтените средства или поне най-малко ”видимо”.

Възприеманият контрол е оценката на индивида доколко е в неговите възможности да осъществи съответното поведение. Макар че на пръв поглед посочването по случаен начин на един от няколко алтернативни отговора не предполага ангажирането на големи или особени по характер поведенчески ресурси, в някои случаи такова поведение може и да не бъде реализирано. Не са редки случаите, при които изпитваните не дават отговор на даден въпрос, на група въпроси или на части от теста, въпреки че биха могли да го направят.

И така, за преобладаващата част от кандидатите може да се предполага, че имат атитюд към поведението („желая да стана студент и ако налучквам, ще имам по-голям шанс да се класирам”), имат готовност за това социално приемливо поведение („всички в залата биха налучквали, когато са затруднени, а и това няма да бъде наказано”) и са в състояние да реализират такова поведение. Мотивацията за високи постижения е свързана с една по-широка потребност на индивида от висока себеоценка, която да подсили положителния му Аз-образ. Подходящ обяснителен модел е социометричната теория, съгласно която себеоценката е показател за това как индивидът се чувства приет от околните (Leary et al., 1995, по Дилова, 2008). Има емпирични данни, че хората с по-висока себеоценка се чувстват по-добре приемани от другите, отколкото тези с ниска себеоценка (ibid.) Най-адекватният механизъм за поддържане и повишаване на висока себеоценка са действията, които носят добри резултати, т. е. високи постижения.

Тук възниква въпросът доколко удовлетворителни за индивида са високите резултати, постигнати (отчасти) с непочтени средства, например висок успех на изпита и влизане в университета, ако определен дял от бала се дължи на налучкване? М. Дилова говори за несъзнавани влияния на потребността от висока себеоценка върху познавателния процес, за определени „изкривявания" на себепознанието, които предпазват индивида от спадане на равнището на себеоценката и на Аз-образа (ibid., стр. 178). В редица експериментални когнитивни изследвания са получени свидетелства за това, че в някои случаи индивидите преработват информацията за себе си по начин, който им позволява да видят себе си в положителна светлина. Хората търсят положителната информация за себе си и избягват отрицателната, склонни са да виждат по-ясно положителните си страни и да ги надценяват, но не и отрицателните(ibid.)

Един подходящ обяснителен механизъм на желанието за налучкване в контекста на потребността от съхраняване и повишаване на себеоценката е феноменът на себеугодното атрибутиране. Той се изразява в стремежа за приписване на успехите на достойнствата собствената личност, а неуспехите- на външни фактори. При провал на изпит от класически тип (с преподавател, застанал пред студента), последният би могъл да припише вината на преподавателя (много е строг, заяжда се, не ме харесва, задава трудни/ провокиращи въпроси и т. н.) или на късмета си (падна ми се въпрос, по който не бях подготвен, говорих след колега, който се представи блестящо и т.н.) При тестов изпит възможностите на студента да атрибутира евентуален неуспех екстернално са твърде ограничени. Един от достъпните начини за решаване на този вътрешен конфликтът е изпитваният да положи допълнителни усилия, тук и сега, по време на изпита, като използва възможността за налучкване, която самата изпитна форма му предоставя.

Третият изследователски въпрос в анализа е свързан със съгласуваността между съответстващите си индекси и параметри, определени в едно и също условие, т.е. при една и съща извадка от и. л.

Бяха направени допусканията, че между стойностите на оценките на трудността на въпросите *p* и *b*, както и между тяхната дискриминативна сила *D* и *a* и r_{bis} и *a*, няма съгласуваност. Това следва от теоретичната вариативност, нестабилност на индексите, определени в рамките на CTT, и тяхната инвариантност в рамките на IRT. Ако даден индекс варира в допустимите граници на неговото изменение, а съответният параметър е стабилен, не би могло да се очаква да има съгласуване между техните стойности.

Резултатите от направените корелационни анализи обаче водят към опровергаване на тези допускания. Наблюдаваните корелационни коефициенти, използвани като мярка за степента на съгласуваност между едноименните статистики, са свидетелство, че между стойностите, получени чрез алгоритмите на двете тестови теории, има определена степен на съответствие. Тук трябва да отбележим видимия контраст между

степента на съгласуваност на статистиките на трудността и на дискриминативната сила на въпросите. Докато между оценките на трудността *p* и *b* се наблюдават, във всички изследвани тестови варианти, изключително високи, статистически значими коефициенти на линейна корелация, то при оценките на дискриминативната сила *D* и *a* и r_{bis} и *a* получените коефициенти на линейна корелация не са така еднопосочни.

Въз основа на получените оценки за степента на съгласуваност между *p* и *b*, на тяхната статистическа значимост, както и на анализа на диаграмите на разсейване на тези статистики можем да заключим, че между оценките на трудността на въпросите, получени в рамките на двете тестови теории, се наблюдава висока степен на съгласуваност, която има ясно изразен линеен характер. С други думи, по отношение на трудността на въпросите Класическата тестова теория и Теорията за отговор на тестов въпрос могат да се разглеждат като взаимнозаменяеми.

Получените резултати подчертават на спецификата на трудността като базова характеристика на тестовите въпроси. Тук следва да обърнем внимание на особеното място, което психометричната общност отделя на еднопараметричния (Раш) модел на IRT, при който единственият вариативен параметър на въпросите е тяхната трудност (*b*). Няма съмнение, че при част от реалните ситуации (без значение дали преобладават или не) този модел не е адекватен на данните – въпросите се третират като притежаващи еднаква (фиксирана) дискриминативна сила (т. е. еднакъв наклон на характеристичната крива), игнорира се и проблемът с налучкването (параметърът *c* също е с фиксирана, нулева стойност). Той обаче притежава едно ценно качество – трудността е единственият параметър, който е разположен на скалата на способностите Θ. Нещо повече, въз основа на вероятностното моделиране на връзката между способността Θ и отговора на съответния въпрос, позицията на този въпрос (неговата трудност като единствена характеристика) може да бъде оценена независимо от това кои лица (т. е. какви индивидуални стойности на Θ) са използвани за тази оценка (Rasch, 2001). Самият Г. Раш, автор на този модел, радващ се на широка популярност, прави образно сравнение на този процес с измерването на температурата на даден обект, което трябва да води към приблизително едни и същи резултати, независимо от това какъв термометър е използван (ibid.)

Както беше отбелязано, различна е картината, която се очертава при изследването на съгласуваността на дискриминативните статистики. Получените оценки са неконсистентни и противоречиви, със сравнително ниски корелационни коефициенти, част от тях – негативни или статистически незначими. Но ако се обърнем към диаграмите на разсейване на съответните двойки статистики, тази неясна ситуация може да намери своето съдържателно обяснение. То се състои в това, че взаимовръзката между статистиките на дискриминативност при всички анализирани тестови варианти има ясно изразен нелинеен характер, който не може да бъде експлициран, поне не толкова адекватно, чрез приложените линейни модели. Изразена чрез корелационното отно-

шение η, при различните варианти степента на нелинейна съгласуваност между *D* и *a* е между 0.50 и 0.70, а степента на нелинейност (η^2 - r^2) варира от 0.30 до 0.50.

При по-детайлен анализ на диаграмите на разсейване на съответните две статистики на дискриминативната сила се очертава още една интересна особеност. При диаграмите на разсейване при почти всички тестови варианти се наблюдава само един ясно изразен локален минимум на апроксимиращата функция. Поради това графиката на съвместното вариране на статистиките на *D* и *a* може да бъде разгледана като съставена от две части, във всяка от които се наблюдава сравнително ясно изразена линейна взаимовръзка. В зоната наляво от локалния минимум на съответната функция корелацията е негативна, а надясно от него - позитивна. Интересно е да се отбележи, че като цяло минималните стойности на параметъра *a* са в областта около 0.20 от *D*, която се приема за долна граница на приемливост на стойностите на този индекс (Ebel, 1954). С други думи, негативната форма на съвместно вариране на *D* и *a* се наблюдава именно в зоната на неприемливите стойности на *D*.

Въз основа на резултатите от направените анализи на съвместното вариране може да се направи заключението, че между статистиките на дискриминативна се наблюдава определена степен на съгласуваност, макар и не така добре изразена, както при статистиките на трудността.

IV. Обща дискусия

Основната изследователска цел в разработката е да се направи сравнително изследване на приложимостта на двете основни психометрични теории, по-конкретно на два модела в техните рамки, върху емпирични данни от Теста за общообразователна подготовка. Концепцията за приложимостта е разгледана в два нейни аспекта: (1) степента на съответствие между допусканията на теоретичния модел и характеристиките на емпиричните тестови данни и (2) степента, в която очакваните свойства на теоретичния модел се проявяват в емпиричните тестови данни, по-конкретно по отношение на очакваното „поведение" на статистиките на тестовите въпроси.

В светлината на предимствата и недостатъците на тези психометрични теории, в изследването са потърсени отговорите на два основни изследователски въпроса:

(1) Съответства ли Класическата тестова теория, в рамките на която функционира ТОП, в достатъчно висока степен на тестовите данни?

(2) Би ли довела замяната на стария психометричен модел с новата Теория за отговор на тестов въпрос, поради нейните безспорни теоретични предимства, до подобряване на измерителните качества на теста?

Изследването е фокусирано върху търсенето на различни по характер свидетелства „за" и „против" приложението на новата психометрична теория при разработването и анализа на резултатите от ТОП. Като отправна точка в него са приети следните два теоретични модела, които съответстват на дизайна на ТОП. В теоретичната рамка на Класическата тестова теория - едномерен, с нормално разпределение на действителния бал, τ-конгенеричен модел. В теоретичната рамка на Теорията за отговор на тестов въпрос - основан на дихотомични отговори, едномерен, с нормално разпределение на латентната способност Θ, параметричен, логистичен модел. В изследването те се разглеждат като „базови модели" със значение на възможни, първоначални рамки за описание на данните от ТОП.

В изследването са поставени и следните две групи от изследователски въпроси, които конкретизират формулираните по-горе общи проблеми. Първата група произтича от разбирането на концепцията за приложимостта като съответствие между допусканията на теоретичния модел и характеристиките на тестови данни:

(1) Каква е размерността на пространството на латентните способности, които обуславят отговорите на и. л. на въпросите от ТОП? Има ли емпирични свидетелства, които да подкрепят допускането за тяхната едномерност?

(2) Каква е формата на разпределенията на латентните способности? Доколко обосновано е допускането, че латентните способности следват нормалното Гаусово

разпределение?

Втората група е свързана с разбирането на приложимостта като проява на очакваните свойства на теоретичния модел в емпиричните тестови данни:

(3) При кой от двата теоретични модела, съответно на CTT и IRT, статистиките на тестовите въпроси са инвариантни в различни условия, т. е. при различни извадки от индивиди?

(4) Доколко статистиките на тестовите въпроси, определени в рамките на един и същи модел, в едно и също условие, т. е. при една и съща извадка от индивиди, функционират независимо една от друга?

(5) Наблюдава ли се съгласуваност между индексите, определени в рамките на CTT, и съответните им параметри, определена в рамките на IRT, в едно и също условие, т. е. при една и съща извадка от индивиди.

Потърсен е и отговорът и на още един въпрос, свързан с базовия модел на IRT:

(6) Дали дефинираният едномерен, с нормално разпределение на латентната способност Θ базов модел на IRT съответства на данните от теста и ако не съответства, кой модел би бил по-подходящ за приложение?

За постигане на изследователските цели са проведени пет относително самостоятелни изследвания, всяко от които има своя дял в изясняването на комплексната проблематика, свързана с приложимостта на психометричните модели. За постигане на пълнота и изчерпателност, изследванията са базирани на няколко методологични подхода. В сърцевината на изследователският анализ на данни, обоснован от Дж. Тюки, стои отказа от „класическия" потвърдителен подход, основан на предварително формулиране на хипотези и последващата им проверка със статистически тестове. Този подход предполага многостранно изучаване на събраните данни чрез системен анализ на взаимовръзките между променливите и прилагане на разнообразни по своя характер методи и средства, чрез натрупване на различни свидетелства за природата на съответствията между теоретичните модели и емпиричните данни.

Вторият подход е свързан с осигуряването на вътрешната валидност на резултатите от изследванията. Методите, използвани в рамките на изследователският анализ на данни, потенциално могат да се отличават с различна степен на адекватност и поради това тяхното използване следва да бъде резултат от обоснован избор. Обикновено понятието „вътрешна валидност" се свързва с изследванията на причинно-следствени отношения. Тук то е използвано в по-тясното му значение – като адекватност на приложената методология, по-конкретно - до адекватния подбор на статистически методи и до коректността на последващата интерпретация на получените чрез тях резултати. Поради това всяко изследване е предшествано от обстоен анализ на качествата на потенциално приложимите алтернативни статистически методи.

Третият подход е свързан с осигуряване на външната валидност на резултатите от изследванията, за възможността тези резултати и направените от тях изводи да бъ-

дат генерализирани върху други съвкупности от тестови данни. От съществено значение е до каква степен направеното изследване е представително за съвкупността от всички възможни изследвания, най-вече от гледна точна на представителността на извадките. В настоящото изследване са обхванати два типа извадки: (1) на изпитаните лица, които представляват част от периодна генерална съвкупност, които могат да бъдат разгледани като случайни и представителни, и (2) на тестовите въпроси, включени в съответния вариант на теста, които не са случайни и представителни. Поради това като мярка за осигуряване на външната валидност е приложен подходът за репликиране на ситуациите, т. е. за анализиране на различни извадки от индивиди и от тестови въпроси чрез паралелно изследване на множество от тестови варианти, използвани в различни точки от време.

Тясно свързан с проблема за външната валидност е и проблемът за осигуряване на екологичната валидност на резултатите. В представеното изследване екологичната валидност на резултатите е обезпечена чрез използването на реални данни, получени от 15 варианта на Теста по общообразователна подготовка, използвани в кандидат-студентските кампании на НБУ през периода 2003 – 2008 година.

Целта на първото емпирично изследване е да се проучи формата на разпределенията на латентните способности и да се установи доколко обосновано е допускането, че тези способности следват нормалната Гаусова крива. Анализите са направени на равнище субтест и цялостен тест. Съгласно концепцията на Дж. Тюки за изследователския подход, при анализа на формата на разпределенията на тестовите балове са приложени четири различни числови и графични статистически методи. Като цяло решенията, които могат да се вземат въз основа на получените резултатите, са съгласувани, макар че при някои конкретни разпределения тези резултати водят и до противоречиви изводи.

Поради неговите предимства, като водещ метод в изследването бе избран статистическият тест на съгласието на Шапиро-Уилк. Резултатите от направените проверки както с този тест, така и с алтернативния тест на Лилиефорс, водят последователно към решения за отхвърляне на нулевите хипотези за нормалност на разпределението на способностите в генерална съвкупност при огромна част от анализираните данни. Анализът на данните чрез останалите числени и графични методи показа различни степени на отдалеченост на техните разпределения от нормалното. Широки са границите на изменение на индексите за асиметрия и ексцес, които при някои от разпределенията достигат екстремни стойности. Отклоненията от стандартните за нормалното разпределение стойности е както в положителна, така и в отрицателна посока. Графичните методи водят към същите заключения. Наблюдават се и малък брой разпределения с екстремно ниски стойности на асиметрия или на ексцес, или едновременно на двата индекса, за които тестовете на съгласието сочат отсъствие на нормалност. Графичните методи, приложени към тези разпределения, също водят към решения за

нормалност.

Доколкото резултатите от тестовете на съгласието се интерпретират дихотомично, при избраното ниво на значимост, а в специализираната литература няма единно мнение относно границите на изменение на индексите за асиметрия и ексцес, за да бъде разглеждана съответната променлива като нормално разпределена, е трудно да се определи каква част от анализираните тестови данни могат да се разглеждат като съответстващи на това допускане. Може да се посочи, че типичното разпределение на тестовия бал се отклонява от нормалното – то е с положителна асиметрия и отрицателен ексцес, а и свидетелствата за нормалност са при сравнително малко на брой данни и произтичат от методи, резултатите от които подлежат на по-субективна интерпретация.

Ето защо, въз основа на приложените различни по характер методи може да се направи обобщението, че разпределенията на тестовите балове в ТОП на равнище субтест и тест следва да се разглеждат като отклоняващи се от нормалното Гаусово разпределение, т. е. разпределението на способностите не е нормално. Оттук следва, че между теоретичното допускане в избраните тестовите модели, по-специално в модела на Теорията за отговор на тестов въпрос, за нормалност на разпределението на латентните променливи, и съответната характеристика на наблюдаваните данни, по правило липсва съответствие. Това несъответствие поставя под въпрос приложимостта на тези теории към данните, получени чрез Теста по общообразователна подготовка. Този извод не следва да се разглежда като изключване на нормалното разпределение като теоретична основа за анализ на тестовите данни. При цялото разнообразие от форми на наблюдаваните емпирични тестови разпределения, Гаусовата крива е тяхната най-подходяща теоретична апроксимация. Разпределенията се различават само по степента на отдалеченост от този теоретичен модел. По-съществените отклонения от нормалността обаче са по-скоро правило, отколкото изключение.

Разгледани в контекста на традицията, завещана от Л. Л. Търстоун, който въвежда нормалното разпределение за моделиране на психологически променливи, а и на практиката в психологическите изследвания, получените резултати изглеждат озадачаващи. Едно възможно обяснение може да бъде потърсено в мнението, че при прилагане на тестовете за нормалност се наблюдава тенденция тестовите статистики да бъдат по-чувствителни при извадки с по-голям обем, каквито са анализираните данни, което води по-често към отхвърляне на нулевата хипотеза. Данните от редица съпоставителни изследвания на различни тестове за нормалност обаче не подкрепят това становище – тестовете имат стабилно поведение при извадки с различен обем. Следователно предположението, че получените в настоящото изследване резултати се дължат на слабост на използваните тестови методи за проверка на нулевите хипотези за нормалност, е неоснователно. От друга страна, макар и малко на брой, публикациите с изследвания върху реални данни показват, че разпределенията с нормална

форма са по-скоро рядък прецедент, отколкото правило - твърде малка част от разпределенията са дори сравнително близка апроксимация на Гаусовото. Резултатите от представеното изследване следва да се разглеждат като още едно свидетелство против тезата за универсалната нормалност.

Далеч по-убедително обяснение на голямото разнообразие от форми на разпределенията, отличаващи се от тази на нормалното, може да се потърси в структурата на отделните извадки. Те са формирани случайно (макар и не в статистическия смисъл на термина) и във всяка от тях са попаднали кандидати, които принадлежат към различни възрастови, социо-културни и демографски групи, с различна образователна история, степен на общообразователна подготовка и мотивация за успех. С други думи, генералната съвкупност се характеризира с определена липса на хомогенност, която се разглежда условие за формиране на нормално разпределение. По-скоро извадките следва да се разглеждат като хетерогенни, съставени от различни по характер субизвадки. Безспорно тази структура намира отражение в тестовите балове, следователно и в техните разпределения.

Не по-малко важна за получените резултати е структурата на способностите, чиито разпределения са обект на изследване. Поради начина на формиране на общия тестов бал като сумарна скала, интегрираща в себе си способностите от отделните раздели, може да се предположи, че той представлява съвместно многомерно разпределение на отделните субтестови способности, което би могло да повлияе силно на формата на неговите разпределения. В предходното изследване беше показано, че подходящ модел за размерността на латентното тестово пространство е тримерният, при наличието на 8-10 първични димензии. От друга страна, макар че някои субтестови разпределения могат да се разглеждат като основно едномерни, незавъртените субтестови конфигурации са също многомерни. Резултатите от изследванията сочат, че, ако пренебрегнем резултатите от тестовете на съгласието, разпределенията на общия тестов бал не се характеризират със системно по-високи отклонения от Гаусовото разпределение в сравнение със субтестовите балове.

И накрая, за отклоненията от нормалната крива принос би могъл да има и един вътрешноприсъщ „дефект” на тестовите балове. Тестовият бал е обобщена мярка на способностите, резултативна величина от използването на дадена сумарна скала, между айтемите на която съществува определена корелация. Парадоксът е в това, че, макар и да са желан ефект от гледна точна на надеждността на скалата, по-високите корелации между айтемите биха довели до разпределения, значително по-плоски от нормалното. Както беше показано по-горе, при разпределенията, анализирани в настоящото изследване, преобладават тези с отрицателен ексцес, който може да бъде обяснен с ефекта на взаимовръзките между отделните (суб)тестови въпроси.

След като тестовете за нормалност на изследваните разпределения на резултатите от ТОП водят до отхвърляне на нулевите хипотези, при достатъчна подкрепа на

тези решения и от другите приложени методи, и при наличието на сравнително голям масив от разпределения (общо 132), върху които е базирано изследването, уместно ли е генерализирането на извода, че разпределенията на способностите се различават от нормалното? Или, ако се върнем към противоположните мнения на К. Хопкинс, Дж. Глас и Р. Гиъри, обича ли Бог нормалната крива или тя е сами мит? Р. Тапиа и Дж. Томпсън поставят въпроса по противоположен начин – възможно ли е да се пренебрегнат такива резултати, получени от ограничен брой разпределения, с ограничен брой наблюдения, значително по-малък от обема на генералната съвкупност?

Възможни са два алтернативни отговора. Първият е, че тези резултати характеризират разпределенията в ограничени по обем извадки и че това не означава непременно, че тяхното разпределение в генералните съвкупности не е нормално. Тази позиция се аргументира с тезата, че с нарастването на обема на извадката разпределението на тестовите балове се стреми към нормалното. Тази теза обаче отразява едно изопачено разбиране на централната гранична теорема, съгласно която към нормалното разпределение се стреми извадковото разпределение на средните стойности на множество такива извадки със същия обем, извлечени от същата генерална съвкупност. При това върху разпределението на наблюденията в генералната съвкупност не се налагат ограничения по отношение на неговата форма. Следователно при наличието на отклоняващи се от нормалното разпределение извадки не може да се направи еднозначният извод, че генералната съвкупност е нормално разпределена.

Алтернативният отговор на поставения въпрос, че ако разпределението на генералната съвкупност не е нормално, е малко вероятно (простите случайни) извадки да придобият нормална форма. Като основно свидетелство можем да посочим медианата на Тюки, съгласно която типичното извадково разпределение не е нормално. Това предположение е по-правдоподобно, защото не противоречи на централната гранична теорема. То се съгласува с огромна част от резултатите от приложените в настоящото изследване методи за оценка на нормалността, както и на данните от литературните източници. Беше показано обаче, че в някои случаи, главно чрез графичните методи, но и при прилагането на консервативните тестове на съгласието, някои разпределения могат да бъдат оценени като нормални. Следователно имаме основания да допуснем, че разпределението в генералната съвкупност е близко до нормалното, но се отклонява от него.

Ако нормалното разпределение е мит, то значи ли това, че Бог го не го обича? Струва ни се,че след всичко казано дотук можем да заключим в същия афористичен стил, че Бог може би е постановил принципа на нормалността, но не се интересува от неговото спазване.

Във второто изследване е направен анализ на размерността на латентните пространства на субтестово и тестово равнище от гледна точка на поставените изследователски въпроси и най-вече доколко обосновано е допускането за тяхната едно-

мерност, отразено в базовите модели.

Незавъртените факторни конфигурации при двете равнища на анализ съдържат множество фактори, чийто брой в повечето от разгледаните случаи надхвърля половината от броя на въпросите в тестовите раздели, съответно в цялостните тестове. В по-голямата си част отделните латентни дименсии имат слабо влияние върху въпросите, не е висока и кумулативната им обяснителна сила. Може да се приеме, следователно, че отговорите на индивидите на тестовите въпроси са обусловени от многомерни латентни структури, които съдържат две групи от различни по своя характер дименсии. Първата група включва общи умствени способности, които влияят върху отговорите на индивидите на всички тестови въпроси. Втората съдържа фактори, които могат да се разглеждат като специфични за всеки тестов въпрос или като проява на случайни процеси. Бяха получени множество свидетелства, че специфичните, уникални фактори играят много по-съществена роля като предиктори на постиженията, отколкото общите фактори.

Докато обичайният подход при определяне на скаловата структура на теста е да се прибегне до някаква стратегия за класифициране на наблюдаваните променливи, за да бъдат удовлетворени изискванията на моделите за едномерност, е необходимо да се постигне, ако има основания за това, някакво „примирие" между тях и многомерните тестови данни. То може да бъде постигнато чрез прилагането на подход, при който се акцентира върху наличието на една ярко изразена латентна характеристика – обстоятелство, което авторите обозначават като наличие на доминиращ фактор, ефективна едномерност или основна едномерност. Според тази концепция, латентното пространство може да се разглежда като едномерно, ако се наблюдава един основен фактор, чрез който може да се обясни голяма част от споделената дисперсия, дори и да е съпроводен от определен брой фактори със слаба обяснителна сила, които се разглеждат като второстепенни. С други думи, пространството може да се разглежда като едномерно, ако отговорите на индивидите на тестовите въпроси са във висока степен функционално зависими от определена доминираща способност. Р. Харви в свое изследване дава пример за такава доминираща дименсия, която обяснява 88% от споделената дисперсия и 51% от цялата дисперсия на айтемите (Harvey, 2003).

На субтестово равнище приложените методи за редуциране на броя на факторите в първоначалните конфигурации не водят до съгласувани и еднозначни и решения. Може да се приеме, че при *част* от анализираните конкретни субтестове съответната латентна структура се характеризира с наличието на един доминиращ фактор, т. е. тези субтестове могат да бъдат разглеждани като основно едномерни и към тях могат да бъдат приложени определените базови модели. В полза на това решение са резултатите от три от приложените четири алтернативни метода. Поради наблюдаваната неустойчивост на факторните конфигурации, при останалата *част* от анализираните субтестове факторната структура е по-скоро аморфна, без наличието на доминиращ

фактор, т. е. с „нулева" факторна структура. Прилагането на базовите, а и на който и да е друг модел, би било безпредметно, доколкото субтестовите резултати не биха могли да се обвържат с конкретна латентна способност. Тази особеност на част от субтестовете е свидетелство за проблеми, свързани с начина, по който са съставени айтемите, както и с конструирането на съответния субтест. В общия случай, всяка субтестова латентна структура попада в една от посочените по-горе две категории. Може да се очаква, че вероятността съответното латентно пространство да бъде едномерно, е по-висока при субтестовете 1. *Български език*, 5. *Математика*, 9. *Разсъждения* и 10. *Семантика*. По-ниска е тази вероятност при субтестове 2. *Литература* и 3. *История*, а най-малко вероятно да се идентифицира едномерно пространство е при субтестове 4. *География*, 6. *Физика*, 7. *Химия* и 8. *Биология*.

На тестово равнище, след неуспешния опит за постигане на проста факторна структура с първични ортогонални фактори и прилагане на йерархичен факторен анализ, латентната структура се очертава като включваща три общи, независими способности от втори ред – за използване на правила, за обобщаване и за възпроизвеждане на знания (паметови способности). Тъй като тези латентни способности бяха идентифицирани при всички анализирани тестове, може да се предположи с увереност, че латентната структура, която лежи в основата на Теста по общообразователна подготовка, е трифакторна. Следователно базовите модели, които предполагат едномерност, са неприложими на равнище тест.

Като обща, характерна особеност на латентните структури на субтестово и тестово равнище може да бъде посочена тяхната обяснителна слабост. На какво се дължи тази слабост? Отговорът на този въпрос може да бъде потърсен в две посоки.

От една страна, в ниска степен на структурираност, на изразеност на съответните латентни способности у самите кандидат-студенти. Причината за това може да бъде потърсена в училищното образование, през което е преминало всяко от и. л., явили се на конкурсен изпит. Изненадващо е, отбелязва Дж. Карол, че независимо от хилядите проучвания върху ефекта на образованието върху развитието на когнитивните способности, не може да се каже, че има системна информация по този въпрос – нито върху общата интелигентност, нито върху специфичните умствени способности. Причината за това е, според автора, че извършването на промяна в умствените способности обикновено не се разглежда като образователна цел (Carrol, 1993). Трудно е да се определи в каква степен образователният процес влияе върху когнитивните способности заради сложната мрежа от взаимовръзки между тях. Авторът привежда резултати от изследвания, които съдържат свидетелства за наличието на каузални връзки между склонността към обучение и академичния успех, които обаче не са еднозначни. Наблюдавана е и позитивна връзка между сложността на учебната програма или нейното съдържание и когнитивните способности.

Мнението, че развиването на умствените способности остава встрани от обра-

зователните цели, е, ако не дълбоко невярно, то поне дискусионно - когнитивните способности следва да се разглеждат като важен фактор за академичните постижения в училище. Тяхното развитие е залегнало в основата на нашата образователна система, ако се съди по Държавните образователни изисквания за учебно съдържание по отделните културно-образователна области. Въпросът, следователно, е в това доколко когнитивните способности се повлияват от обучението и дали образователната система прилага в достатъчна степен средствата и методите, чрез които да съдейства активно или поне да благоприятства за тяхното развитие. Отговорът на първия въпрос е по-скоро положителен, ако се съди по резултатите от някои изследвания. Показателна в това отношение е новата генерация на теста SAT, който по своето предназначение и дизайн е сходен на ТОП. След 2005 г. SAT е много по-тясно обвързан с учебните програми и дисциплини в средното училище (SAT subject tests). Най-същественото изменение в идеологията на теста е отказът, по препоръка на Р. Аткинсън, от използването на SAT Reasoning test, за който се приема, че измерва „вродени" (*innate/ native*) способности, като основен критерий за прием в университетите. Новият SAT е инструмент за оценка на умения за критично мислене и разсъждения, които се развиват през времето (*developed ability*) чрез опита, който индивидите придобиват във и извън училище (Atkinson, 2001; Zwick, 2007). Отговорът на втория въпрос е по-скоро отрицателен. Съдейки по резултатите от направените анализи, нашата образователна система не успява да се справи с тази задача. Системното развитие на способностите е по-скоро декларирана цел, която стои на заден план пред усвояването на знания за факти, събития, обстоятелства и явления.

От друга страна, причината може да бъде потърсена в дизайна на самия инструмент за оценяване, който не е предназначен за разкриване на идентифицираните структури от латентни способности. На субтестово равнище, както беше отбелязано, той е предназначен за оценка на знания и умения, обвързани с отделни предметни области – езикови, литературни, математически и т. н. Установените в изследването общи фактори обаче са „надпредметни", необвързани с отделните субтестове. Всичко това поставя под съмнение валидността на ТОП, на субтестово и тестово равнище, по отношение на неговата конструктна валидност.

Следващите три изследвания са посветени на друг важен аспект на концепцията за приложимостта, който се изразява в степента, в която очакваните свойства на двете психометрични теории се проявяват в емпиричните тестови данни, най-вече по отношение на очакваното „поведение" на статистиките на тестовите въпроси. Анализът на тестовите данни доведе до множество интересни, в някои случаи изненадващи резултати, които не се съгласуват с направените предположения. Очертаха се и някои важни тенденции, които хвърлят нова светлина върху съвместното поведение на тестовите статистики.

Първият изследователски въпрос в тази част на анализа е свързан със стабил-

ността, инвариантността на статистиките на въпросите, оценени в рамките на двете психометрични теории. За оценка на този аспект от тяхното поведение бяха приложени два взаимно допълващи се подхода – корелационен анализ за изследване на относителната съгласуваност на стойностите на съответния индекс или параметър, получени в две различни условия, и дисперсионен анализ с повторни измервания – за оценка на съотношението между централните им тенденции при първото и второто измерване.

Като цяло получените коефициенти на стабилност при всеки от наблюдаваните индекси и параметри, определени в рамките двете теории, се характеризират с много високи, статистически значими равнища. Тези резултати могат да се разглеждат като свидетелство, че както индексите, определени в рамките на *CTT*, така и параметрите, определени в рамките на *IRT*, се отличават с висока степен на устойчивост, на стабилност по отношение на относителните им позиции при последователно оценяване в различни условия. Следва да се отбележи, че коефициентите на стабилност на всеки индекс или параметър, независимо върху коя двойка от тестови варианти са определени, се характеризират с близки, съпоставими стойности, което е още едно свидетелство за тяхната устойчивост и възпроизводимост.

В рамките на *CTT* системно по-високи са оценките на стабилността на индекса за трудност (*p*) на въпросите, който може да се разглежда като тяхна най-устойчивата характеристика. По-малко устойчив е индексът на дискриминативна сила, а най-малко – бисериалният коефициент на корелация. В рамките на *IRT* се наблюдава градация на коефициентите на стабилност, подобна на предходната. И тук със системно по-високи равнища се отличава стабилността на параметъра на трудност (*b*) на въпросите, малко по-ниски са равнищата при параметрите на дискриминативна сила (*a*) и на налучкване на правилния отговор (*c*). Следователно може да се заключи, че двете тестови теории, приложени върху данните от ТОП, са в еднакво годни да осигурят висока степен на устойчивост на относителните равнища на статистиките на въпросите, оценени в различни условия.

Не са така сходни резултатите, които произтичат от прилагането на втория критерий за устойчивост на статистиките на въпросите. Поставянето на тестовите въпроси в различни условия, при различни извадки от и. л., не предизвиква статистически значими разлики в средните равнища на техните индекси, определени в рамките на *CTT*. Обратно, при параметрите, определени в рамките на IRT, единствено параметърът на трудността (*b*) демонстрира стабилност. При другите два от тях – дискриминативна сила (*a*) и налучкване на правилните отговори (*c*), се наблюдава статистически значимо изместване на средните стойности. Следователно, по отношение на втория критерий двете тестови теории, приложени върху данните от ТОП, не са равностойни. Прилагането на алгоритмите на *CTT* води по-често и при повече статистики на въпросите до по-устойчиви резултати, отколкото тези на новата психометрична теория.

Ако приложим възприетия конюнктивен подход, съгласно който поведението на

статистиките следва да удовлетворява едновременно двата критерия за инвариантност, следва да направим обобщението, че допускане 1 (а) за зависимост на индексите на трудност (*p*) и на дискриминативна сила (*D,* r_{bis}), определени в рамките на *CTT*, от извадките, въз основа на които са получени, не се потвърждава. Не намира опора в реалните данни и допускане 1 (б) за независимост на параметрите на дискриминативна сила (*a*), трудност (*b*) и налучкване на правилния отговор (*c*), определени в рамките на *IRT*, от извадките, въз основа на които са получени. С други думи, статистиките на въпросите, определени в рамките на CTT, са по-устойчиви на промени в условията, отколкото тези, определени в рамките на IRT.

Наблюдаваната инвариантност на статистиките на въпросите в рамките на *CTT* не следва да се разглежда като проява на тяхна независимост. Какъв е психометричният смисъл на получените резултати? Доколкото индексът на трудност (*p*) отразява относителния дял на правилните отговори на даден въпрос при дадена извадка от и. л., неговата стабилност следва да се интерпретира като свидетелство, че този относителен дял се съхранява и възпроизвежда в различни тестови сесии, при различни групи от и. л. Дискриминативният индекс отразява и друга особеност на извадката – нейната хомогенност по отношение на измервания признак. Възпроизвеждането на равнищата на този индекс при различни извадки може да се обясни със съхраняването на приблизително една и съща структура на извадките по отношение на този признак. С други думи, устойчивостта на индексите е обусловена от относително стабилната структура на извадките по отношение на измерваните способности.

Очакваната стабилност на параметрите на въпросите по *IRT*, тяхната независимост от извадката от и. л. бе подложена на съмнение, независимо от това, че в теоретичен план параметрите са характеристики на самите айтеми, но не и на групата, въз основа на която са изчислени. Един от източниците на вариативност следва да се потърси в чувствителността на характеристичната крива на въпросите към промени в относителния дял на правилните отговори в различни точки от скалата, които не биха се отразили на класическите индекси. Други източници на вариативността следва да бъдат потърсени в особеностите на популационното разпределение. Както бе показано, анализът на неговата форма и размерност водят към заключението, че способностите не са разпределени нормално и не формират едномерно разпределение. А това са две от основните допускания, които формират фундамента на разглеждания модел на *IRT*. Тук може да се добави и възможното негативно влияние на обемите на извадките, които вероятно не са достатъчно големи, както и адекватността на логистичната характеристична крива към всеки отделен въпрос.

Следва да се отбележи, че числовите стойности на параметрите, получени в хода на анализа на тестовите данни, не представляват действителните им стойности, а са техни оценки. Тяхната вариативност е свидетелство за наличието на отклонения на наблюдаваните от действителните стойности. Тя не е основание да се подложат на

съмнение теоретичните достойнства на изследвания модел на новата психометрична теория. Вариативността им обаче е свидетелство за негативния ефект от несъответствието между теоретичния модел и реалните данни.

Вторият изследователски въпрос в анализа е свързан с взаимовръзките между разноименните индекси, от една страна, и параметри, от друга, определени въз основа на данните от отделни тестови варианти. Бяха направени две допускания за очакваното поведение на статистиките: (а) за наличие на нелинейна взаимовръзка между стойностите на индексите на трудност (*p*) и на дискриминативна сила (*D*), определени в рамките на *CTT* върху една и съща извадка и 2 (б) за липса на взаимовръзки от корелационен или функционален тип между стойностите на параметрите на дискриминативна сила (*a*), трудност (*b*) и налучкване на правилния отговор (*c*), определени в рамките на *IRT*.

Резултатите от направените анализи свидетелстват, че между всички статистики, определени в рамките на дадена теория, се наблюдават ясно изразени взаимовръзки от корелационен тип.

Особен интерес представлява взаимовръзката между индексите на въпросите в рамките на *CTT* и по-конкретно между индексите на дискриминативна сила *D* и трудност *p*. Между двата индекса бе установена криволинейна връзка, която има форма на изпъкнала парабола. Наблюдават се очакваните ниски абсолютни стойности на дискриминативния индекс в зоните на екстремно високи и ниски стойности на трудността, както и високи стойности на дискриминативния индекс в средната част на скалата на трудността. Съвместното вариране на индексите беше апроксимирано по метода на нелинейната регресия с полиномна функция от втора степен. Беше определено и съответното регресионно уравнение с удовлетворителни оценки на неговите параметри. В диаграмата на разсейването се съдържат нагледни свидетелства за това, че изпитваните прилагат стратегия за налучкване на правилния отговор и че тази стратегия е по-присъща на лицата от „слабата", отколкото на тези от „силната" група.

Следователно може да се направи обобщението, че допускането за наличие на нелинейна връзка между индексите на трудността *p* и дискриминативна сила *D*, определени в рамките на *CTT*, намира опора в емпиричните данни. Разбира се, наличието на такава взаимовръзка следва да се разглежда като недостатък, като „присъща" слабост на Класическата теория.

Неочаквани се оказаха резултатите от изследването на взаимовръзките между параметрите на тестовите въпроси, определени в рамките на *IRT*. Тук от особен интерес е взаимовръзката между дискриминативната сила (*a*) и трудността (*b*), която може да бъде разгледана не само от гледна точка на очакването за липса на връзка между двата параметъра, но и от тази на демонстрираната криволинейна връзка между аналогичните статистики в *CTT*.

Анализът на диаграмите на разсейване показа, че подходящ модел на взаимов-

ръзката е нелинейният, който може да бъде представен чрез полиномна функция от четвърта степен. Особеност на сложната крива е това, че нейният нелинеен характер се проявява най-вече в двата края на континуума, в зоните на високите (отрицателни или положителни) стойности на параметъра *b*. Не по-малко интересна е наблюдаваната взаимовръзка между параметъра за налучкване на правилния отговор (*c*) и останалите два параметъра. При всички тестови варианти тя е негативна, статистически значима, с високи стойности, които при анализите на взаимовръзката на параметъра на налучкване с дискриминативната сила са системно по-високи. Взаимовръзката между тези параметри може да се разглежда като линейна, макар че при всички диаграми на разсейване се наблюдава слабо изразена нелинейност.

Следователно може да се направи заключението, че предположението за липса на взаимовръзка между параметрите на въпросите, определени чрез алгоритмите на *IRT*, не намира опора в емпиричните данни.

Нека да обърнем внимание на обстоятелството, че параметърът на налучкване на правилния отговор е негативно свързан с дискриминативната сила и трудността на въпросите. С други думи, с увеличаване на стойността на първия параметър се намаляват стойностите на останалите два. На какво се дължи наблюдаваната висока негативна корелация?

Известно е, че корелационните отношения не маркират непременно каузална връзка между съответните променливи. В този случай обаче допускането на еднопосочно влияние на параметъра на налучкване върху другите два параметъра би било добра основа за разбирането на наблюдавания ефект.

Включването на параметъра на налучкване в описанието на тестовите въпроси отразява една житейска реалност – индивидите, които се явяват на изпит, могат да посочат правилния отговор, дори и когато нямат необходимите знания и умения. Чрез този параметър в зависимата променлива - вероятността за правилен отговор, се включва приносът на този феномен.

Съгласно трипараметричния модел на *IRT*, параметърът на налучкване е постоянен във всички точки на континуума на способностите, за него също е валиден принципът за независимост от извадката. Дали обаче този параметър е характеристика, която произтича само от въпроса? Действително, някои особености на въпросите влияят пряко върху този параметър. Но налучкването следва да се разглежда и като поведение, характеризиращо даден индивид или група индивиди. Ако застанем на тази позиция, връзката между параметъра на налучкване и другите два параметъра става по-прозрачна и може да получи правдоподобно обяснение.

Ако при отговора на даден въпрос определена група от лица прибегне до стратегията на налучкване, по-вероятно е това да не са всички лица от извадката, а само тези с по-ниски способности, които се намират на левия край на континуума на способностите. Свидетелство за това бяха получени при анализа на взаимовръзката между

индексите в рамките на *CTT*. В резултат от прилагането на тази стратегия, върху лявата част на характеристичната крива ще бъде оказан „натиск" отдолу нагоре, който би повишил долната й граница. Повишаването на вероятността от правилен отговор в лявата част на скалата, без да е съпроводено с подобно повишаване в дясната й част, предполага промяна в разпределението на и. л. на скалата на способностите, която се изразява в отслабване на дискриминативната сила на въпроса. Това е механизмът, за който може да се предположи, че регулира съвместната вариация на двата параметъра.

Наличието на статистически значими негативни корелации между параметрите на налучкване и трудността на въпросите означава, че има тенденция въпросите да намаляват трудността си с увеличаване на вероятността от налучкване, т. е. да изместват позицията си наляво на скалата. Този феномен може да бъде обяснен с нарастването на общата маса на правилните отговори.

Може да се предположи, следователно, че силата, която отключва тази мрежа от взаимодействия, е налучкването на правилните отговори като съзнателно поведение на част от и. л. В този смисъл параметърът на налучкване може да бъде разглеждан, наред със способностите, като втори параметър в *IRT*, който съдържа личностов компонент и който отразява определени психични процеси. Такова поведение на индивидите може да бъде разгледано в светлината на теориите за мотивацията. Без съмнение, то е обусловено от мотивацията за високи постижения, която включва множество потребности и мотиви за действие, насочени към постигането на високи резултати и значими цели.

Класическият модел на мотивационния процес "очаквания – ценност" на Х. Хекхаузен представя избора на дадено поведение като последица от очакването, че това поведение ще доведе до определен резултат, и от субективната желателност или ценност на очаквания резултат. В изпитни ситуации като тези, при които са събрани анализираните данни, са налице и двата компонента. Кандидатстудентските изпити са с висок залог, имат състезателен характер и положителният изход от тяхното провеждане (преминаване в друг статус) е желан и има висока стойност за кандидатите. В този смисъл прибягването до стратегията за налучкване, чрез която даден кандидат може да придобие определено предимство и да надхвърли действителния си бал, може да се разглежда като мотивирано от очакването, това поведение би подпомогнало постигането на желания резултат.

И. Айзен и М. Фишбайн разработват по-детайлна теория на подбудителната регулация (за планираното поведение), в освен фактора „атитюд към поведението" (очаквания – ценност), въвеждат още два фактора – „субективна норма" и „възприеман контрол".

Като цяло у нас външната среда демонстрира нетърпимост към използването на непочтени методи за постигане на определено предимство. Но в една изпитна ситуа-

ция социалната норма се установява по-скоро от участниците в изпита, която толерира използването на такива средства като преписване и подсказване, дори и при изпити, които имат съревнователен характер. Нормата за приемливо поведение може да бъде зададена и от конкретен авторитет – лицето (институцията), която провежда изпита. Обикновено изпитващите не поставят такива ограничения, дори напротив, при изпълнението на много тестови програми изпитваните са окуражавани да посочат какъвто и да е отговор на въпросите, по които не се чувстват уверени. Макар че при провеждане на Теста по общообразователна подготовка кандидат-студентите не са явно поощрявани да налучкват, няма ограничение, което да ги възпира. Възприеманият контрол е оценката на индивида доколко е в неговите възможности да осъществи съответното поведение. Доколкото посочването по случаен начин на един от няколко алтернативни отговора не предполага ангажирането на големи или особени по характер поведенчески ресурси, и. л. са в състояние да реализират такова поведение.

Мотивацията за високи постижения е свързана с една по-широка потребност на индивида от висока себеоценка, която да подсили положителния му Аз-образ. Съгласно социометричната теория, себеоценката е показател за това как индивидът се чувства приет от околните. Има емпирични данни, че хората с по-висока себеоценка се чувстват по-добре приемани от другите, отколкото тези с ниска себеоценка. Най-адекватният механизъм за поддържане и повишаване на висока себеоценка са действията, които носят добри резултати, т. е. високи постижения.

Последното изследване е посветено на съгласуваността между съответстващите си индекси и параметри, определени в едно и също условие, т.е. при една и съща извадка от и. л. Бяха направени допусканията, че между стойностите на оценките на трудността на въпросите *p* и *b*, както и между тяхната дискриминативна сила *D* и *a* и r_{bis} и *a*, няма съгласуваност. Тези предположения следват от теоретичната вариативност, нестабилност на индексите, определени в рамките на *CTT*, и тяхната инвариантност в рамките на *IRT*.

Резултатите от направените корелационни анализи обаче водят към отхвърляне на тези допускания. Наблюдаваните коефициенти, използвани като мярка за степента на съгласуваност между едноименните статистики, са свидетелство, че между стойностите, получени чрез алгоритмите на двете тестови теории, има определена степен на съответствие. Може да се отбележи контрастът между степента на съгласуваност между статистиките на трудността и на дискриминативната сила на въпросите.

Резултатите от използваните числови и графични методи свидетелстват, че между оценките на трудността на въпросите, получени в рамките на двете тестови теории, се наблюдава висока степен на съгласуваност, която има ясно изразен линеен характер. С други думи, по отношение на трудността на въпросите CTT и IRT могат да се разглеждат като взаимнозаменяеми. Тези резултати подчертават спецификата на трудността като базова характеристика на тестовите въпроси. Тук следва да обърнем

внимание на особеното място, което психометричната общност отделя на еднопараметричния (Раш) модел на *IRT*, при който единственият вариативен параметър на въпросите е тяхната трудност. Този модел притежава едно ценно качество – трудността е единственият параметър, който е разположен на скалата на способностите и неговата стойност може да бъде оценена независимо от това кои лица (т. е. какви индивидуални стойности на Θ) са използвани за тази оценка.

При анализа на съгласуваността на дискриминативните статистики беше установена силно изразена, нелинейна корелация. При по-детайлен анализ на диаграмите на разсейване на съответните две статистики на дискриминативната сила се очертава още една интересна особеност. При диаграмите на разсейване се наблюдава само един ясно изразен локален минимум на апроксимиращата функция. Поради това графиката на съвместното вариране на статистиките може да бъде разгледана като съставена от две части, във всяка от които се наблюдава сравнително ясно изразена линейна взаимовръзка. В зоната наляво от локалния минимум на съответната функция корелацията е негативна, а надясно от него – позитивна, като негативната форма на съвместно вариране се наблюдава в зоната на неприемливите стойности на класическия дискриминативен индекс. Въз основа на резултатите от направените анализи на съвместното вариране може да се направи заключението, че между статистиките на дискриминативна се наблюдава определена степен на съгласуваност, макар и не така добре изразена, както при статистиките на трудността.

Доколкото Тестът по общообразователна подготовка се разработва в рамките на Класическата тестова теория, която от теоретична гледна точка страда от редица недостатъци, основната цел на разработката бе да се отговори на въпроса доколко тази теория (чрез определения базов модел) съответства на тестовите данни. Вторият основен изследователски въпрос бе дали преминаването от класическия модел към Теорията за отговор на тестов въпрос би довело, поради нейните безспорни теоретични предимства, до подобряване на измерителните качества на теста.

Съпоставянето на резултатите от направените анализи дава основание да се заключи, че между „поведението" на двата теоретични модела, приложени върху широка съвкупност от тестови данни, има редица очевидни сходства.

Съгласно резултатите от изследванията на основните допускания на моделите, тестовите данни се характеризират с различна степен на отдалеченост от нормалната Гаусова крива. Използването на тестовите модели би довело до определени неточности в оценките на статистиките, базирани на това разпределение. И тъй като изискването за нормалност е императивно за IRT, при която всички статистики са основани на нормалното разпределение, нейното приложение би било по-малко обосновано от това на CTT.

Тестовите данни се характеризират със слабо изразена, многофакторна латентна структура от общи способности. При част от анализираните субтестове тя може да

се приеме за основно едномерна и към такива данни могат да се приложат и двата модела. На тестово равнище, въпреки наличието на доминиращ първи фактор и множество многофакторни въпроси, латентната структура може да се разглежда като тримерна. Това поставя под въпрос използването на който и да е от двата модела.

Съгласно резултатите, представени в разработката, стабилността, инвариантността на индексите на въпросите, определени в рамките на CTT, е не само напълно съпоставима с тази на параметрите по IRT, но и в много отношения я превъзхожда. Това е може би най-малко очакваната разлика в поведението на двете тестови теории. Опровергано беше и още едно допускане – между параметрите на въпросите, определени в рамките на IRT, съществуват силни взаимовръзки така, както съществуват и между индексите по CTT. Двете теоретични рамки са сходни и поради това, че между едноименните статистиките на въпросите има висока степен на съгласуваност, особено по отношение на трудността.

Тъй като двата тестови модела бяха последователно приложени към една и съща съвкупност от тестови данни, сравнително еднаквото им представяне следва да се разглежда като свидетелство, че на приеманото за даденост теоретично превъзходство на Теорията за отговор на тестов въпрос не се е осъществило. Данните от Теста по общообразователна подготовка са среда, към която „меката" и по-„непретенциозна" Класическата тестова теория се приспособява по-добре, отколкото новата психометрична теория. Като основна причина за това можем да посочим несъответствието между реалните тестови данни, които се отличават с определена степен на неподреденост и неорганизираност, на аморфност, и изискванията на далеч по-сложната в концептуално, структурно и математическо отношение Теория за отговор на тестов въпрос. Следователно преминаването от класическия модел към Теорията за отговор на тестов въпрос като рамка за разработването на Теста по общообразователна подготовка в сегашния му вид, не би довело да съществено подобряване на неговите измерителни качества.

IRT като теоретична концепция съществува под формата на множество модели, съдържащи различен набор от допускания, предназначени за удовлетворяване на различни видове данни. Въз основа на резултатите от изследванията бихме могли да очертаем профила на модела, който би бил по-подходящ за приложение върху Теста по общообразователна подготовка. Този модел следва да бъде освободен от допускане за нормалност на разпределенията на променливите, каквито са моделите на Рамзи, непараметричните модели, както и тези, основани на асиметрични разпределения. За анализ на данните на тестово равнище по-подходящ би бил многомерен, в частност – тримерен, трипараметричен модел. За съжаление, тези модели са по-сложни и не така добре разработени както „стандартния" модел, послужил като основа на направените изследвания.

Трябва да се отбележи, че в представените изследвания бяха засегнати далеч не всички аспекти на приложимостта. Необходимо е да се извърши немалка по обем изследователска работа за изясняване на проблемите, свързани с локалната независимост на отговорите; с диференциалното функциониране на отделните айтеми и на теста; с адекватността на логистичния модел на характеристичната крива на въпросите; със съгласуваността между наблюдавания бал по CTT и оценките на способностите по IRT; с надеждността и валидността на резултатите на субтестово и тестово равнище, включително конструктната и предиктивната валидност; с изясняване на латентната структура на субтестово и тестово равнище и по-прецизното характеризиране на латентните способности. Трябва да се отбележи, че търсенето на адекватен тестов модел не трябва да се разглежда като самоцел. За да изпълнява добре своите функции, Тестът по общообразователна подготовка следва да отговаря на четири основни характеристики:

- валидност на резултатите – особено конструктната и предиктивната валидност;
- надеждност на резултатите;
- ефективност – изпитните процедури не следва да отнемат повече ресурси (материални, човешки, финансови), отколкото е необходимо;
- приемливост – всички заинтересовани страни да се доверяват на резултатите и да бъдат удовлетворени от тяхното качество.

Изпълнението на тази нелека задача вероятно би могло да бъде постигнато чрез търсенето на приемлив компромис между подходящ, работещ теоретичен модел и промяна в дизайна на теста, който да съответства на получените емпирични резултати.

ЦИТИРАНА ЛИТЕРАТУРА

1. Анастази, А., Урбина, С. (2001). *Психологическое тестирование*. Санкт-Петербург: Питер.
2. Величков, А. (1989). *Личност и вътрешна мотивация*. С.: Издателство на БАН.
3. Герганов, Е. (1976*). Психометрични методи за проверка и оценка на знания по български език*. С.:, Народна просвета.
4. Гласс, Дж, Стэнли, Дж. (1976). *Статистические методы в педагогике и психологии*. Москва: Прогресс.
5. Дилова, М. (2008). *Експериментална психология на себепознанието*. С.: Нов български университет.
6. Зиновиева, И. (2009) Екологичен подход към изследване на индивидуалността. В: *Годишник на Софийския университет „Св. Климент Охридски, Философски факултет, книга Психология*, т. 102.
7. Калинов, К. (2002). *Практическа статистика за археолози и антрополози*. С.: Нов български университет.
8. Калинов, К. (2010). *Статистически методи в поведенческите и социалните науки*. С., Нов български университет.
9. Лазарсфельд, П. (1973). Латентно-структурный анализ и теория тестов. В: *Математические методы в социальных науках*. М.: Прогресс, стр. 42-53.
10. Стоименова, Е. (2000). *Измерителни качества на тестовете*. С., НБУ
11. Суппес, П., Зинес, Дж. (1967) Основы теории измерений. В: *Психологические измерения*. Под ред. Л. Д. Мешалкина. Москва: Мир, сс. 9-110.
12. Abedi, J. (1996). The interrater/ test reliability system (ITRS). *Multivariate Behavioral Research*, 31, 4, 409-417.
13. Abelson, R. P., Tukey, J. W. (1959). Efficient conversion of non-metric into metric information. *Proceedings of the Social Statistics Section of the American Statistical Association*. Washington, pp. 226-230.
14. Adedoyin, O., Nenty, H. & 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.
15. Aiken, L. R. (1988). *Psychological Testing and Assessment*. Massachusetts: Allyn & Bacon, Inc.
16. Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, pp. 179-211
17. Allen, S. J., Hubbard, R. (1986). Regression equations for the latent roots of random data correlation matrices with unities on the diagonal. *Multivariate Behavioral Research*, 21, pp. 393-398.
18. Amarnani, R. (2009). Two theories, one theta: A rentle Introduction to Item pesponse theory as an alternative to Classical test theory. *The International Journal of Educational and Psychological Assessment*, Vol. 3, pp. 104-109.
19. 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.
20. Assessment Systems Corporation (1997). *User's Manual for XCALIBRE™ for Windows: Marginal Maximum-Likelihood Estimation Program*. St. Paul MN: Author.
21. Atkinson, R. (2001, February 18). *Standardized tests and access to American universities*. The 2001 Robert H. Atwell Distinguished Lecture, delivered at the 83rd annual meeting of the American Council on Education, Washington, D.C.
22. Baker, B. O., Hardyck, C. D., Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' prescriptions on statistics. *Educational and Psychological Measurement*, 26, pp. 291-309.
23. Baker, F. B. (2001). *The basics of Item response theory*. ERIC Clearinghouse on Assessment and Evaluation, 2-nd ed.

Извлечено на 5.12.2006 от: http://info.worldbank.org/etools/docs/library/117765/Item%20Response%20Theory%20-%20F%20Baker.pdf
24. Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology*, Statistical section, 3, pp. 77-85.
25. Barton, M. A., Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, N.J.: Educational Testing Service.
26. Bazan, J., Bolfarine, H. & Branco, M. (2004). *A new family of asymmetric models for item response theory: A Skew-Normal IRT family*. Извлечено на 15.11.2010 от: http://argos.pucp.edu.pe/~jlbazan/download/paperAPMfinal.pdf
27. 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.
28. Behrens, J. (1997). Principles and procedures of Exploratory data analysis. *Psychological methods*, Vol. 2, No. 2, pp. 131-160.
29. Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In: F. M. Lord and M. R. Novick (eds). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, pp. 395-479.
30. Bloom, B. S., Mesia, B. B., & Krathwohl, D. R. (1964). *Taxonomy of Educational Objectives: The classification of educational goals* (two vols: The Cognitive domain & The Affective domain &). New York: David McKay Co., Inc.
31. Bock, D., Moore, E. (1986). *Advantage and disadvantage: a profile of American youth*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.
32. Bock, R. D., Mislevy, R. J. (1981) An item response curve model for matrix-sampling data: The California grade 3 assessment. In: D. Carlson (ed.), *Testing in the States: beyond accountability*. S.F.: Jossey-Bass.
33. Bolt, D. M., Cohen, A. S. & Wollack, J. A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, 26(4), pp. 381-409.
34. Boyle, J., Fisher, S. (2007). Educational testing. A competence-based approach. Text for the British psychological society's Certificate of competence in educational testing (Level A). Oxford, The British Psychological Society and Blackwell Publishing Ltd. Извлечено на 19.11.2010 от: library.wur.nl/WebQuery/catalog/lang/1861197
35. Bradley, J. W. (1980). Nonrobustness in z, t, and F tests at large sample sizes. *Bulletin of the Psychonomic Society*, 16, pp.333-336.
36. Breckler, S. J. (1990). Application of covariance structure modelling in psychology: Cause for concern? *Psychological Bulletin*, 107, pp. 260-273.
37. Brennan, R. L. (2001). *Generalizability theory*. NY: Springer-Verlag.
38. Brodin, U., Fors, U. & Laksov, K. (2010). The application of Item response theory on a teaching strategy profile questionnaire. *BMC Medical Education*, 10:14 Извлечено на 27.04.2012 от: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2830224/
39. Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.
40. Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkley: University of California Press.
41. Buja, A., Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27, pp. 509-540.
42. Cardinet, J., Tourneur, Y. & Allal, L. (1976). The symmetry of Generalizability theory: Applications to educational measurement. *Journal of Educational Measurement*, Vol. 13, No. 2, pp. 119-135.
43. Carrol, J. B. (1993). Human cognitive abilities: A survey of factor-analytic studies. Ca.: Cambridge University Press.
44. Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, 1, pp. 245-276.
45. Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences*. NY: Plenum Press.
46. Cattell, R. B., Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research,* 12, pp. 289–325.
47. Chen, L., Shapiro, S. (1995). An alternative test for normality based on normalized spacings.

Journal of Statistical Computation and Simulation, 53, pp. 269-287.
48. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
49. Cohen, R. J, Swerdlik, M. E. (2005). *Psychological testing and assessment: An Introduction to Tests and Measurement*. New York: McGraw-Hill, 6th ed.
50. Colton, D. A., Gao, X., Harris, D. J., Kolen, M. J., Martinovich-Barhite, D., Wang, T. & Welch, C. (1997). Reliability issues with performance assessments: A Collection of Papers. *ACT Research Report*, Series 97-3.
51. Cook, T. D., Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Chicago, Illinois: Rand McNally College Pub. Co.
52. Cooke, D. J, Michie. C., Hart, S. D. & Hare, R. D. (1999). Evaluating the screening version of the Hare Psychopathy Checklist – Revised (PCL: V): An item response theory analysis. *Psychological Assessment*, Vol 11(1), pp. 3-13.
53. Coombs, C. H. (1950). Psychological scaling without a unit of measurement. *Psychological Review*, 57, pp. 145-158.
54. Coombs, C. H. (1964). *A theory of data*. New York: John Wiley and Sons, Inc.
55. Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, 78 (1), pp. 98-104.
56. Costello, A. B., Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, 10 (7), pp. 1-9.
57. Courvoisier, D., Eid, M. & Nussbeck, F. (2007). Mixture Distribution State-Trait-Models: Basic ideas and applications. *Psychological Methods, 12*, 80-104.
58. Crocker, L., Algina, J. (1986*). Introduction to classical and modern test theory*. New York: Holt, Rinehart, and Winston.
59. Cronbach, L. J. (1951) Coefficient alpha and the internal structure of tests. *Psychometrika*, 16 (3), 297-334.
60. Cronbach, L., Gleser, G., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons, Inc.
61. D'Agostino, R. B. (1971) An omnibus test of normality for moderate and large size samples. *Biometrika*, 58, pp. 341-348.
62. D'Agostino, R. B., Belanger, A., & D'Agostino Jr., R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, 44, pp. 316-322.
63. Darlington, R. B. (1977). *Factor analysis*.
Извлечено на 24.09.2011 от: http://www.psych.cornell.edu/darlington/factor.htm
64. Dawson, T. E. (2003). *Basic concepts in classical test theory: Relating variance partitioning in substantive analyses to the same process in measurement analyses*. Texas A&M University.
Извлечено на 14.12.2003 от: http://www.tamu.edu/
65. de Ayala, R. J. (2009). *The theory and application of Item response theory*. N.Y.: Guilford Publishing.
66. DeVellis, R. F. (2003). Scale development: theory and applications. *Applied Social Research Methods Series*, Vol. 26. Thousand Oaks, Ca.: Sage Publications, Inc., 2-nd ed.
67. Downing, S. M., Haladyna, T. M. (eds.) (2006). *Handbook of test development*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
68. Ebel, R. L. (1954). Procedures for the analysis of classroom tests. *Educational and Psychological Measurement*, 14 (2), 352-364.
69. 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.
70. Eid, M. (1996). Longitudinal confirmatory factor analysis for polytomous item responses: model definition and model selection on the basis of stochastic measurement theory. *Methods of Psychological Research Online*, Vol. 1, No. 4.
Извлечено на 26.03.2013 от: http://www.pabst-publishers.de/mpr/
71. Elashoff, J. D., Elashoff, R. M. (1978). Effects of errors in statistical assumptions. In: W. H. Kruskal and J. M. Tanur (eds.), *International encyclopedia of statistics*, New York: Free Press, pp. 229-250.
72. Embretson, S. E., Hershberger, S. L. (eds.) (1999). *The new rules of measurement: What*

every educator and psychologist should know. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
73. Embretson, S. E., Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
74. Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 3, pp. 272-299.
75. Fagot, R. F. (1959). A model for ordered metric scale by comparison of intervals. *Psychometrica*, Vol. 24, No. 2, pp. 157-168.
76. 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.
77. Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.
78. Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The applications of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel psychology*, 39, pp. 291-314.
79. Frigg, R., Hartmann, S. (Winter 2006 Edition). Models in science. *The Stanford Encyclopedia of Philosoph.* Edward N. Zalta (ed.)
Извлечено на 16.03.2010 от: http://plato.stanford.edu/entries/models-science/
80. Gaito, J. (1960). Scale classification and statistics. *Psychological Review*, 67, pp. 277-278.
81. Gardner, P. L. (1975). Scales and statistics. *Review of Educational Research*, Vol. 45, No. 1, pp. 43-57.
82. Geary, R. C. (1947). Testing for normality. *Biometrika*, 34, No. 3/ 4, pp. 209-242.
Извлечено на 6.01.2012 от:
http://webspace.ship.edu/pgmarr/Geo441/Readings/Geary%201947%20-%20Testing%20for%20Normality.pdf
83. George, D., Mallery, P. (2003). *SPSS for Windows step by step: A simple guide and reference*. 11.0 update (4-th ed.). Boston: Allyn & Bacon.
84. Gibson, J. J. (1979). *An ecological approach to visual perception*. Boston: Houghton Mifflin.
85. Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement*, 55, pp. 377-393.
86. Golub, G. H., Van Loan, C. F. (1983). *Matrix computations*. Johns Hopkins University Press.
87. Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: What they are and how to use them. *Educational and Psychological Measurement*, Vol. 66, No. 6, pp. 930-944.
88. Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
89. Gribbons, B., Herman, J. (1997). True and quasi-experimental designs. *Practical Assessment, Research & Evaluation*, 5(14).
90. Gronlund, N. E., Linn, R. L. (1990). *Measurement and evaluation in teaching*. New York, Macmillan, 6th ed.
91. Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, 10 (4), 255-282.
92. Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, 19, pp. 149-162.
93. Hakstian, A. R., Muller, V. J. (1973). Some notes on the number of factors problem. *Multivariate Behavioral Research*, Vol. 8 (4), pp. 461-475.
94. Hambleton, R. K. (1989). Principles and selected applications of Item response theory. In R. L. Linn (ed.), *Educational measurement*, 3-rd ed., pp. 147-200. N. Y.: Macmillan.
95. Hambleton, R. K. (Ed.) (2000). Advances in performance assessment methodology. *Applied Psychological Measurement*, 24(4), pp. 291-378.
96. 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.
97. Hambleton, R. K., Rovinelli, R. J. (1986). Assessing the dimensionality of a set of test items. *Applied psychological measurement*, 10, 287-302.
98. Hambleton, R. K., Swaminathan, & H., Rogers, H. J. (1991). *Fundamentals of Item response*

theory. Newbury Park, Ca.: Sage Publications, Inc.
99. Hambleton, R. K., Swaminathan, H. (1984). *Item Response Theory: Principles and Applications*. Hingham, MA: Kluwer, Nijhoff.
100. Hambleton, R. K., Swaminathan, H. (1985). *Item response theory*. Boston: Kluwer-Nijhoff.
101. Harman, H. H. (1976). *Modern factor analysis*. Chicago: The University of Chicago Press.
102. Harman, H. H., Jones, W. H. (1966). Factor analysis by minimizing residuals (minres). *Psychometrika*, 31 (3), pp. 351-368.
103. Harris, D. (1993). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, 12 (3), pp. 157-163.
104. Harris, R. J. (2001). *A primer of multivariate statistics*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
105. Hartwig, F., Dearing, B. E. (1979). Exploratory data analysis. *Sage University Papers Series on Qualitative Research Methods*, Vol. 16, Newbury Park, Ca.: Sage Publications, Inc.
106. Harvey, R. J. (2003, april). 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.
107. Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in Exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, 7:191, pp. 191-205.
108. Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, 58, pp. 78-79.
109. 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.
110. Hill, M., Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics*, 38, pp. 377-396.
111. Hoaglin, D. C., Mosteller, F. & Tukey, J. W. (eds.) (1991). *Fundamentals of exploratory analysis of variance*. John Wiley & Sons, Inc.
112. Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *American Statistical Association Journal*, 69, pp. 909-927.
113. Holland, P., Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly nonparallel test. *Psychometrika*, Vol. 68, No. 1, pp. 123-149.
114. Holt, E.W, Cook, E. F., Covar, R. A., Spahn, J., & Fuhbrigge, A. L. (2008). Identifying the components of asthma health status in children with mild to moderate asthma. *Journal of Allergy and Clinical Immunology*,121, pp. 1175-1180.
115. Hopkins, K. D., Glass, G. V. (1978). *Basic statistics for the behavioral sciences*. Englewood Cliffs, NJ: Prentice-Hall.
116. Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, pp. 179-185.
117. 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.
118. Hulin, C. L., Drasgow, F. & Parsons, C. K. (1983). Item response theory: Applications to psychological measurement. Homewood, IL.: Dow Jones-Irwin.
119. Hurley, A. E., Scandura, T. A., Schriesheim, C. A., Brannick, M. T., Seers, A., & Vandenberg, R. J. (1997). Exploratory and confirmatory factor analysis: Guidelines, issues, and alternatives. *Journal of Organizational Behavior*, 18, pp. 667-683.
120. Jolliffe, I. T. (2002). *Principal component analysis*. NY: Springer-Verlag, Inc.
121. Jöreskog, K. G. (1966). Testing a simple stucture hypothesis in factor analysis. *Psychometrika*, Vol. 31, No. 2, pp. 165-178.
122. Kabacoff, R. I. (2003*). Determining the dimensionality of data: A SAS® macro for Parallel analysis*.
Извлечено на 15.09.2011 от: http://www.mrg.com/articles/parallel.sas.
123. Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, pp. 141-151.
124. Kasleman, H. J., Huberty, C., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L. Petoskey, M. D. & Kaselman, J. C. (1998). Statistical practices of edu-

cational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research*, 68, pp. 350-386.
125. Keeling, K. B (2000). A regression equation for determining the dimensionality of data. *Multivariate Behavioral Research*, 35, pp. 457-468.
126. Kellaghan, T., Greaney, V. (2001). *Using assessment to improve the quality of education*. Paris, UNESCO, International Institute for Educational Planning. Извлечено на 14.06.2009 от: http:// www.unesco.org/iiep
127. Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology*, 30, pp. 17-24.
128. 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.
129. Kim, J. O. (1975). Factor Analysis. In: N. H. Nie, C. H. Hall, S. G. Jenkins, K. Steinbrenner and D. H. Brent (Eds.) *SPSS: Statistical package for the social sciences*. NY: McGraw Hill, 2nd ed.
130. Kim, J. O., Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*. Newbury park, Ca.: Sage Publications, Inc.
131. 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. Извлечено на 21.09.2010 от: http://eric.ed.gov/PDFS/ED268141.pdf
132. Kline, T. J. (2005). *Psychological testing: a practical approach to design and evaluation*. Thousand Oaks, Ca.: Sage Publications, Inc.
133. Knol, D. L., Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, 26, pp. 457-477.
134. Kubinger, K. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, Vol. 45, (1), pp. 106-110.
135. Kuder, G. F., Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2, pp. 151−160.
136. Labovitz, S. (1967). Some observations on measurement and statistics. *Social Forces*, 46, pp. 151-160.
137. Lane, D. (2007). *Online statistics education: A multimedia course of study*. Извлечено на 06.09.2011 от: http://onlinestatbook.com/
138. Lawley, D. N. (1943, January). On problems connected with item selection and test construction. In: *Proceedings of the royal society of Edinburg. Section A. Mathematical and Physical Sciences*, Vol. 61, Issue 03, pp. 273-287
139. Lazarsfeld, P. (1960) Evidence and inference in social research. In: D. Lerner (ed.) *Evidence and inference*. N.Y.: Free Press.
140. Lazarsfeld, P. (1969). A conceptual introduction to latent structure analysis. In: P. F. Lazarsfeld (ed.). *Mathematical thinking in the social sciences*. Glencoe: Free Press, pp. 349-387.
141. Lazarsfeld, P. F., Henry, N. W. (1968). *Latent structure analysis*. Boston: Houghton Mifflin.
142. Leary, M. R. Tambor, E. S., Terdal, S. K. & Downs, D. L. (1995). Self-esteem as an interpersonal monitor: The sociometer hypothesis. *Journal of Personality and Sosial Psychology*, 68, pp. 518-530.
143. Ledesma, R. D., Valero-Mora, P. (2007). Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out Parallel analysis. *Practical Assessment, Research & Evaluation*, Vol. 12, 2. Извлечено на 30.09.2009 от: http://pareonline.net/getvn.asp?v=12&n=2
144. 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. Извлечено на 12.08.2012 от: http://www.aare.edu.au/03pap/lee03219.pdf
145. 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, pp. 114-126.
146. Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown, *Journal of American Statistical Association*, Vol. 62, pp. 399-402.

147. Linacre, J. M. (2004). Item discrimination, guessing and carelessness: Estimating IRT parameters with Rasch. *Rasch Measurement Transactions*, Vol.18, No. 1.
148. Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement*, 13, pp. 517-548.
149. Lord, F. M. (1980). *Applications of Item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
150. Lord, F. M., Novick, M. R. (1974). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.
151. Lyerly, S. B. (1958). The Kuder-Richardson formula (21) as a split-half coefficient, and some remarks on its basic assumption. *Psychometrika*, Vol. 23, Number 3, pp. 267-270.
152. 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.
153. McCabe, G. P. (1978). *Use of the 27% rule in experimental design*. Purdue university, Department of statistics, Division of mathematical sciences, Mimeograph series, No. 449 Извлечено на 14.03.2011 от: www.stat.purdue.edu/research/technical_reports/.../tr-499.pdf - United States
154. McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
155. McNemar, Q. (1969). *Psychological statistics* (4th ed.) N.Y.: John Wiley & Sons, Inc.
156. Micceri, T. (1989). The Unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, Vol. 105. No.1, pp. 156-166.
157. Micceri, T. (2003). *A discussion of statistical robustness*. Извлечно на 11.05.2011 от: http://www.freewebs.com/tedstats/Files/Robustness.pdf
158. Michell, J. (1999). *Measurement in psychology: critical history of methodological concept*. Cambridge: Cambridge University Press.
159. Miller, K., Ramaswami, S., Rousseeuw, P., Sellares, J., Souvaine, D., Streinu, I., & Struyf, A. (2003). Efficient computation of location depth contours by methods of computational geometry. *Statistics and Computing*, Vol. 13, 2, pp. 153-162.
160. Miller, M. B. (1995). Coefficient alpha: A basic introduction from the perspectives of classical test theory and structural equation modeling. *Structural Equation Modeling*, 2, pp. 255-273.
161. Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics*, Vol. 8, No. 4, pp. 271-288.
162. Mislevy, R. J. (1984). Estimating latent populations. *Psychometrika*, 49, pp. 359-381.
163. Molin, P., Abdi, H. (1998). *New tables and numerical approximation for the Kolmogorov-Smirnov/ Lillierfors/ Van Soest test of normality*. Technical report, University of Bourgogne. Извлечено на 14.10.2010 от: http://www.utd.edu/herve/MolinAbdi1998-LillieforsTechReport.pdf
164. 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.
165. Nandakumar, R. (1993). Assessing essential unidimensionality of real data*. Applied Psychological Measurement*, Vol. 17, No. 1, pp. 29-38.
166. Nandakumar, R., Ackerman, T. (2004). Test modeling. In: D. Kaplan (ed.), *The Sage handbook of quantitative methodology in the social sciences*, pp. 93-105. Thousand Oaks, CA: Sage Publications.
167. Nandakumar, R., Yu, F., Li, H. & Stout, W. (1998) Assessing unidimensionality of polytomous data. *Applied Psychological Measurement*, Vol. 22, pp.99-115.
168. 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.
169. Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
170. O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, 32, pp. 396-402. Извлечно на 26.10.2012 от: http://people.ok.ubc.ca/brioconn/nfactors/nfactors.html
171. Park, H. M. (2008). *Univariate Analysis and Normality Test Using SAS, Stata, and SPSS*. Working Paper. The University Information Technology Services (UITS), Center for Statistical

and Mathematical Computing, Indiana University.
Извлечено на 29.01.2011 от: http://www.indiana.edu/~statmath/stat/all/normality/index.html
172. Pedhazur, E. J., Schmelkin, L. P. (1991). *Measurement, design and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
173. Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis*. Thousand Oaks, Ca.: Sage Publications, Inc.
174. Phillips, J. P. (1971). A note on the representation of ordered metric scaling. *British Journal of Mathematical and Statistical Psychology*, Vol. 24, Issue 2, pp. 239-250.
175. Pollard, B., Dixon, D., Dieppe, P., & Johnston, M. (2009). Measuring the ICF components of impairment, activity limitation and participation restriction: an item analysis using classical test theory and item response theory. *Health and Quality of Life Outcomes*, 7 (41). Извлечено на 16.05.2011 от: http://www.hqlo.com/content/7/1/41
176. Popham, W. J. (1981). *Criterion-referenced measurement*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.
177. Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
178. Rasch, G. (2001). On objectivity and specificity of the probabilistic basis for testing. In: *Rasch Lectures. In honor of Georg Rasch's 100 years birthday on the 21th of September, 2001*. Eds. L. Olsen and S. Kreiner. Copenhagen Business School.
извлечено на 20.12.2012 от: http://www.rasch.org/memos.htm#Georg
179. Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research*, No. 32, pp. 329-353
180. Raykov, T. (1998). A method for obtaining standard erros and confidence intervals of composite reliability for congeneric items. *Applied Psychological Measurement*, 22 (4), pp. 369-374.
181. Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. *Journal of Educational Statistics*, Vol. 4, No. 3, pp. 207-230
182. Reckase, M. D. (1990, april). *Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests*. Paper, presented at the annual meetion of the American educational research association, Boston, MA.
183. 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.
184. Revelle, W. (2007). *Determining the number of factors: the example of the NEO-PI-R*. Department of Psychology, Northwestern University
Извлечено на 07.04.2011 от: http://www.personality-project.org/r/book/numberoffactors.pdf
185. Revelle, W. (2011*). An overview of the psych package*. Department of Psychology, Northwestern University.
Извлечено на 07.04.2011 от: http://personality-project.org/r/overview.pdf
186. Revelle, W., Rocklin, T. (1979). Very simple structure - alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, 14 (4), pp. 403–414.
187. Reynolds, C., Kamphaus, R. (eds) (2003). *Handbook of psychological & Educational assessment of children*. New York, NY: The Guilford Press.
188. Riekert, K. A., Eakin, M. (2008). Factor analysis: A primer for asthma researchers. *Journal of Allergy and Clinical Immunology*, 121, pp. 1181-1183.
189. Ripley, B. D. (2004). *Robust statistics*.
Извлечено на 22.04.2011 от: http://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf
190. Rosenbaum, P. R. (1988). Item bundles. *Psychometrika*, Vol. 53, pp. 349-359.
191. Rousseeuw, P. J., Ruts, I. and Tukey, J. W. (1999). The bagplot: A bivariate boxplot. *The American Statistician*, 53 (4), pp. 382–387.
192. Rousseeuw. P. J., Ruts, I. (1996). Bivariate location depth. *Applied Statistics*, 45, pp. 516-526.
193. Rousseeuw. P. J., Ruts, I. (1998). Constructing the bivariate Tukey median. *Statistica Sinica*, 8, pp. 827-839.
194. Royston, J. P. (1982) An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, No. 2, pp. 115-124.
195. Royston, J. P. (1986). A remark on AS181: The W Test for normality. *Applied Statistics*, 35, pp. 232-234.

196. Royston, P. (1989). Correcting the Shapiro Wilk W for ties. *Journal of Computation and Simulation*, 31, pp. 237-249.
197. Royston, P. (1992) Approximating the Shapiro-Wilk W-test for non-normality. *Statistics and Computing*, 2, pp. 117-119.
198. Ryan, T. A., Joiner, B. L. (1976). *Normal probability plots and tests for normality*. Statistics Department, The Pennsylvania State University.
Извлечено на 05.05.2011 от:
www.minitab.com/uploadedFiles/.../normal_probability_plots.pdf
199. Samejima, F. (1995). Acceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, 60, pp. 549–572.
200. Samejima, F. (1997). Departure from normal assumptions: a promise for future psychometrics with substantive mathematical modeling. *Psychometrika*, 62, 4, pp. 471-493.
201. Samejima, F. (2000). Logistic positive exponent family of models: Virtue of asymmetric item characteristic curves. *Psychometrika*, 65, pp. 319-335.
202. Sax, G. (1989). *Principles of educational and psychological measurements and evaluation* (3rd ed.) Belmont, Ca.: Wadsworth.
203. Seier, E. (2002). *Comparison of tests for univariate normality*.
Извлечено на 15.10.2010 от:
http://interstat.statjournals.net/YEAR/2002/articles/0201001.pdf
204. Shapiro, S. S., Wilk, M. B. (1965). An analysis of variance test for normality (Complete samples). *Biometrika*, Vol. 52, No. 3/4, pp. 591-611.
205. Shapiro, S. S., Wilk, M. B. and Chen, H. J. (1968). A comparative study of various tests for normality. *Journal of the American Statistical Association*, No. 63, pp. 1343–1372.
206. Shavelson, R. J., Webb, N. M. (1991). *Generalizability theory: A primer*. Ca.: Sage Publications, Inc.
207. Siegel, S. (1956). *Non-parametric statistics for the behavioural sciences*. NY: McGraw-Hill.
208. Sočan, G. (2000). Assessment of reliability when test Items are not essentially τ-equivalent. Developments in survey methodology. In: A. Ferligoj and A. Mrvar (eds), *Metodološki zvezki*, 15, Ljubljana: FDV
Извлечено на 07.05.2008 от: http://ams.sisplet.org/uploadi/editor/mz15socan.pdf
209. Spearman, C. (1904). “General intelligence”, objectively determined and measured. *American Journal of Psychology*, Vol. 15, No. 2, pp. 201-293.
Извлечено на 14.11.2010 от: http://psychclassics.yorku.ca/Spearman/
210. Steiger, J. H. (1994). Factor Analysis in the 1980’s and the 1990’s: Some old debates and some new developments. In: Ingwer Borg and Peter Ph. Mohler (Eds.), *Trends and perspectives in empirical social research*. Berlin: Walter de Gruyter.
211. Steiger, J. H. (2009). *Measures of fit in structural equation modeling: An introduction*.
Извлечено на 26.10.2012 от:
http://www.statpower.net/Content/312/Handout/Measures%20of%20Fit%20in%20Structural%20Equation%20Modeling.pdf
212. Steiger, J. H., Lind, J. C. (1980). *Statistically-based tests for the number of common factors*. Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City.
Извлечено на 16.12.2012 от: http://www.statpower.net/Steiger%20Biblio/Steiger-Lind%201980.pdf
213. Steiger, J., Shapiro, A. & Browne, M. (1985). On the multivariate asymptotic distribution of sequential chi-square statistics. *Psychometrika*, Vol. 50, No. 3. pp. 253-264.
214. Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, Vol. 69, No. 347, pp. 730-737.
215. Stevens, J. (2002). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
216. Stevens, S. S. (1939). On the problem of scales for the measurement of psychological magnitudes. *Journal for Unified Science*, 1939, Vol. 9, pp. 94-99.
217. Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, Vol. 103, No. 2684, pp. 677-680.
218. Steyer, R. (2001). *Classical (Psychometric) test theory*. Friedrich-Schiller-Universitaet Jena, Institut fuer Psychologie, Lehrstuhl fuer Methodenlehre und Evaluationsforschung.
Извлечено на 17.12.2009 от: www.metheval.uni-jena.de/materialien/publikationen/ctt.pdf
219. Steyer, R., Ferring, D., & Schmitt, M. (1992). States and traits in psychological assessment.

European Journal of Psychological Assessment, 8, pp. 79-98.
220. Steyer, R., Majcen, A., Schwenkmezger, P. & Buchner, A. (1989). A latent state-trait anxiety model and its application to determine consistency and specificity coefficients. *Anxiety Research*, 1, pp. 281-299.
221. Steyer, R., Schmitt, M. & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13, 389–408.
222. Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press.
223. Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrica*, Vol. 52, pp. 589-617.
224. Stryjewski, L. (2011). *40 years of boxplots*. Извлечено на 13.04.2011 от: http://vita.had.co.nz/papers/boxplots.pdf
225. Suppes, P. (1962). Models of data. In: E. Nagel, P. Suppes & A. Tarski (eds.) *Logic, methodology and philosophy of science: Proceedings of the 1960 International congress*. Stanford: Stanford University Press, pp. 252-261.
226. Tapia, R. A., Thompson, J. R. (1978). *Nonparametric probability density estimation*. Baltimore, MD: Johns Hopkins University Press.
227. Tate, R. L. (1995). Robustness of the school-level IRT model. *Journal of Educational Measurement*, Vol. 32, No. 2, pp 145-162.
228. Taylor, J. M. (1985). Measures of location of skew distributions obtained through Box-Cox transformations. *Journal of the American Statistical Association*, 80, pp. 427-432.
229. Theune, J. A. (1973). Comparison of power for the D'Agostino and the Wilk-Shapiro test of normality for small and moderate samples. *Statistica Neerlandica*, Vol. 27, Issue 4, pp. 163-168.
230. Thissen, D., Steinberg, L. & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, Vol. 26, pp. 247-260.
231. Thompson, B. (2003). Guidelines for authors, reporting score reliability estimates. In: B. Thompson (ed*.). Score reliability: Contemporary thinking on reliability issues*. Ca.: Sage Publications, Inc.
232. Thorndike, R. M. (1971). *Metod of factor extraction and simple structure of data from diverse scientific areas.* Paper presented at the annual meeting of the Western psychological association, San Francisco, CA. Извлечено на 23.02.2009 от: http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED056075&ERICExtSearch_SearchType_0=no&accno=ED056075
233. Thorndike, R. M., Cunningham, G. K., Thorndike, R. L. & Hagen, E. P. (1991*). Measurement and evaluation in psychology and education*. (5th ed.) New York: Macmillan.
234. Thurstone, L. L. (1927). A Law of comparative judgment. *Psychological Review*, 34, pp. 273-286. Извлечено на 15.05.2011 от: http://mpg.ndlab.net/wp-content/uploads/2009/07/thurstone94law.pdf
235. Thurstone, L. L. (1929). The measurement of psychological value. In: T. V. Smith and W. K. Wright (eds.). *Essays in philosophy by seventeen doctors of philosophy of the University of Chicago*. Chicago: Open Court, pp. 157-174. Извлечено на 01.09.2011 от: http://www.brocku.ca/MeadProject/Thurstone/Thurstone_1929a.html
236. Thurstone, L. L. (1931a). *The reliability and validity of tests*. Ann Arbor, MI: Edwards Brothers.
237. Thurstone, L. L. (1931b). Multiple Factor Analysis. *Psychological Review*, 38, pp. 406–427. Извлечено на 24.01.2012 от: http://psychclassics.yorku.ca/Thurstone/
238. Thurstone, L. L. (1934). The vectors of mind. *Psychological Review*, 41, pp. 1-32. Извлечено на 24.01.2012 от: http://psychclassics.yorku.ca/Thurstone/
239. Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. Chicago: University of Chicago Press. Извлечено на 24.01.2012 от: http://psychclassics.yorku.ca/Thurstone/
240. Thurstone, L. L. (1936). The factorial isolation of primary abilities. *Psychometrica*, 1, pp. 175–182.
241. Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Извлечено на 24.01.2012 от: http://psychclassics.yorku.ca/Thurstone/
242. Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: John Wiley & Sons, Inc.
243. Tukey, J. W. (1975). Mathematics and the picturing of data. In: *Proc. of the 1974 International congress of mathematicians*, Vancouver, pp. 523-531.
244. Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
245. Uebersax, J. S. (2000). *Estimating a latent trait model by factor analysis of tetrachoric correlations*.
Извлечено на 10.05.2011 от: http://john-uebersax.com/stat/irt.htm
246. Vehkalahti, K. (2000). Reliability of measurement scales. *Statistical Research Reports*, 17. Helsinki: The Finnish Statistical Society.
Извлечено на 18.02.2010 от: http://ethesis.helsinki.fi/
247. Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41, pp. 321-327
248. Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In: R. D. Goffin and E. Helmes (eds.) *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy*. Norwell, MA: Kluwer Academic.
249. Wainer, H., Bradlow, E., & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.
250. Wainer, H., Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, 15, pp. 22-29.
251. Wang, X., Wainer, H., Brown, L., Bradlow, E. Skorupski, W., Boulet, J., & Mislevy, R. (2006). An application of Testlet response theory in the scoring a complex certification exam. In: D. Williamson, R. Mislevy, & I. Bejar (eds). *Automated scoring of complex tasks in Computer based testing*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., pp. 169-199.
252. Weiner, I. B., Freedheim, D. K., Schinka, J. A., & Velicer, W. F. (2003). *Handbook of Psychology: Research methods in psychology*. NJ: John Wiley & Sons, Inc.
253. Weng, L. J., Cheng, C. P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement*, Vol. 65, No. 5. pp. 697-716.
254. 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).
Извлечено на 30.03.2006 от: http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-12629
255. Wilcox, R. R., Charlin, V. L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics*, 11, pp. 263-274.
256. Wilkinson, L., & APA Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.
257. Wolf, D., Bixby, J., Glenn, J. & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (ed.), *Review of Research in Education*, 17, pp. 31-125.
258. Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for non-normal latent variables. *Psychological Methods*, Vol. 11(3), 253-270.
259. Woods, C. M. (2007). Ramsay-curve IRT for Likert-type data. *Applied Psychological Measurement*, Vol. 31, No. 3, 195-212.
260. Woods, C. M. (2008) Ramsay-curve Item response theory for the three-parameter logistic Item response model. *Applied Psychological Measurement*, Vol. 32, No. 6, 447-465.
261. Wright, B. D. (1999). Fundamental measurement for psychology. In: S. E. Embretson & S. L. Hershberger (eds.) *The new rules of measurement: What every educator and psychologist should know,* pp. 65-104. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
262. Wright, B. D., Mead, R.J. (1976). BICAL: Calibrating items with the Rasch model. *Research Memorandum No. 23*. Statistical Laboratory, Department of Education, University of Chicago.
263. Wright, B. D., Stone, M. H. (1999). *Measurement essentials*. Wilmington, DE: Wide Range, Inc.
Извлечено на 17.06.2010 от: http://www.rasch.org/memos.htm#measess
264. Yates, A. (1987). *Multivariate exploratory data analysis: a perspective on exploratory factor analysis*. Albany: State university of New York press.

265. 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.
266. Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8, pp. 338-353.
267. Zimmerman, D. W. (1976) Test theory with minimal assumptions. *Educational and Psychological Measurement*, 36, pp. 85-96.
268. Zwick, W. R. (2007, february). *College Admission Testing*. A report commissioned by the National Association for College Admission Counseling.
Извлечено на 24.01.2013 от: http://education.ucsb.edu/rzwick/FinalNACAC.pdf
269. Zwick, W. R., Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, 99, pp. 432-442.

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ 1.

Хронология на администрирането на ТОП

Календарна година	Брой изпитни дати	Брой варианти на ТОП	Брой изпитани
1996	6	6	1 635
1997	2	4	3 357
1998	5	8	9 343
1999	13	43	10 048
2000	6	22	9 351
2001	6	22	10 818
2002	6	24	9 762
2003	6	18	10 170
2004	7	29	11 222
2005	6	20	12 134
2006	6	20	13 692
2007	6	20	14 394
2008	6	20	16 218
2009	9	30	14 674
2010	10	28	9 395
общо	100	314	156 213

ПРИЛОЖЕНИЕ 2.

Варианти на ТОП, подбрани за анализ

Вариант на ТОП	Брой изпитани	Изпитна дата
Вар. 092	636	20.04.2003
Вар. 096	652	18.04.2004
Вар. 110	865	27.06.2003
Вар. 127	721	25.07.2004
Вар. 128	543	27.02.2005
Вар. 132	638	17.04.2005
Вар. 134	713	17.04.2005
Вар. 141	730	23.07.2005
Вар. 146	454	24.07.2005
Вар. 154	764	16.04.2006
Вар. 166	835	23.07.2006
Вар. 171	830	10.06.2007
Вар. 175	925	29.04.2007
Вар. 192	1019	13.04.2008
Вар. 198	934	08.06.2008
	общ брой: 11 259	

ПРИЛОЖЕНИЕ 3.

Резултати от проверката на хипотезите за нормалност на разпределенията на тестовите балове на равнище субтест и тест

Вариант на ТОП	Раздел	Колмогоров-Смирнов *D*	Lillefors *p*	Shapiro-Wilk's *W*	Shapiro-Wilk's *p*	Асиметрия	Станд. грешка на асиметрията	Ексцес	Станд. грешка на ексцеса
1	2	3	4	5	6	7	8	9	10
Вар. 92	Total	0.056	<0.01	0.988	0.00	0.338	0.097	-0.203	0.194
	1. Български	0.140	<0.01	0.962	0.00	-0.154	0.097	-0.174	0.194
	2. Литература	0.111	<0.01	0.972	0.00	0.225	0.097	-0.308	0.194
	3. История	0.151	<0.01	0.953	0.00	0.353	0.097	0.007	0.194
	4. География	0.138	<0.01	0.965	0.00	0.166	0.097	-0.251	0.194
	5. Математика	0.135	<0.01	0.953	0.00	0.544	0.097	-0.134	0.194
	6. Физика	0.169	<0.01	0.954	0.00	0.346	0.097	-0.197	0.194
	7. Химия	0.161	<0.01	0.945	0.00	0.167	0.097	-0.252	0.194
	8. Биология	0.134	<0.01	0.963	0.00	0.284	0.097	-0.009	0.194
	9. Разсъждения	0.128	<0.01	0.966	0.00	0.202	0.097	-0.615	0.194
	10. Семантика	0.172	<0.01	0.944	0.00	-0.418	0.097	-0.042	0.194
Вар. 96	Total	0.068	<0.01	0.984	0.00	-0.043	0.096	1.311	0.191
	1. Български	0.126	<0.01	0.965	0.00	0.076	0.096	-0.067	0.191
	2. Литература	0.128	<0.01	0.966	0.00	0.287	0.096	-0.313	0.191
	3. История	0.142	<0.01	0.967	0.00	0.139	0.096	-0.194	0.191
	4. География	0.131	<0.01	0.965	0.00	0.317	0.096	-0.035	0.191
	5. Математика	0.146	<0.01	0.960	0.00	0.472	0.096	0.188	0.191
	6. Физика	0.162	<0.01	0.953	0.00	0.251	0.096	-0.180	0.191
	7. Химия	0.155	<0.01	0.943	0.00	0.519	0.096	0.675	0.191
	8. Биология	0.150	<0.01	0.950	0.00	0.246	0.096	0.462	0.191
	9. Разсъждения	0.114	<0.01	0.956	0.00	-0.509	0.096	0.311	0.191
	10. Семантика	0.130	<0.01	0.956	0.00	-0.310	0.096	0.745	0.191
Вар. 110	Total	0.068	<0.01	0.989	0.00	0.386	0.083	0.353	0.166
	1. Български	0.160	<0.01	0.947	0.00	0.530	0.083	0.006	0.166
	2. Литература	0.123	<0.01	0.965	0.00	0.313	0.083	-0.220	0.166
	3. История	0.128	<0.01	0.969	0.00	0.127	0.083	-0.173	0.166
	4. География	0.114	<0.01	0.969	0.00	0.019	0.083	-0.251	0.166
	5. Математика	0.158	<0.01	0.948	0.00	0.650	0.083	0.535	0.166
	6. Физика	0.136	<0.01	0.961	0.00	0.249	0.083	-0.020	0.166
	7. Химия	0.179	<0.01	0.925	0.00	0.559	0.083	0.455	0.166
	8. Биология	0.145	<0.01	0.958	0.00	-0.014	0.083	0.226	0.166
	9. Разсъждения	0.115	<0.01	0.970	0.00	-0.084	0.083	-0.593	0.166

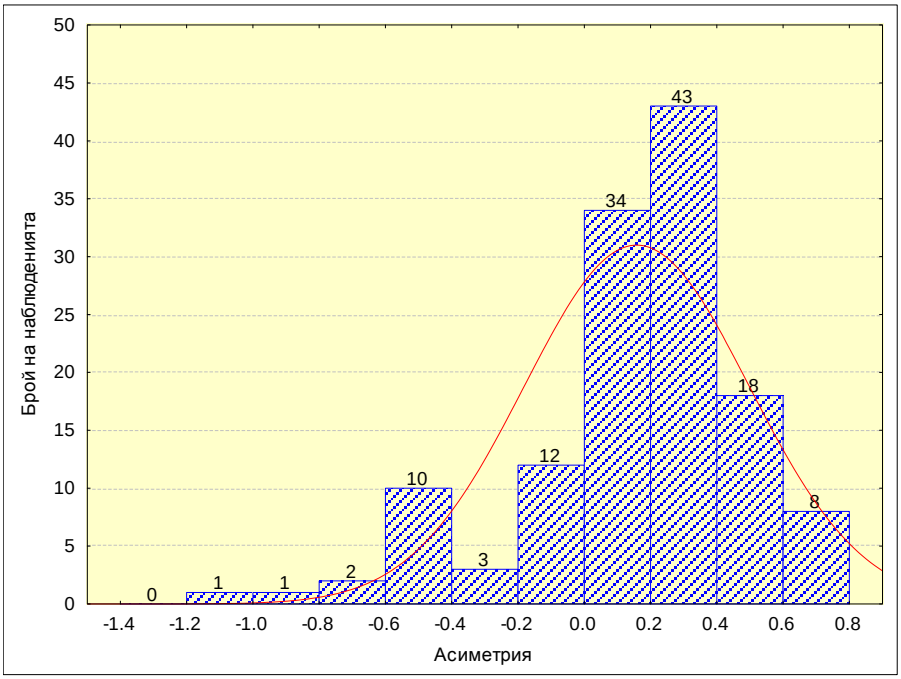
Вариант на ТОП	Раздел	Колмогоров-Смирнов *D*	Lillefors *p*	Shapiro-Wilk's *W*	Shapiro-Wilk's *p*	Асиметрия	Станд. грешка на асиметрията	Ексцес	Станд. грешка на ексцеса
1	2	3	4	5	6	7	8	9	10
	10. Семантика	0.182	<0.01	0.926	0.00	-0.816	0.083	1.198	0.166
Вар. 127	Total	0.059	<0.01	0.990	0.00	0.357	0.091	0.068	0.182
	1. Български	0.132	<0.01	0.965	0.00	0.275	0.091	-0.274	0.182
	2. Литература	0.118	<0.01	0.969	0.00	0.181	0.091	-0.446	0.182
	3. История	0.125	<0.01	0.968	0.00	0.025	0.091	-0.228	0.182
	4. География	0.159	<0.01	0.942	0.00	0.584	0.091	0.381	0.182
	5. Математика	0.154	<0.01	0.935	0.00	0.783	0.091	0.965	0.182
	6. Физика	0.160	<0.01	0.955	0.00	0.268	0.091	-0.131	0.182
	7. Химия	0.169	<0.01	0.936	0.00	0.439	0.091	0.041	0.182
	8. Биология	0.125	<0.01	0.969	0.00	0.186	0.091	-0.218	0.182
	9. Разсъждения	0.125	<0.01	0.966	0.00	0.171	0.091	-0.083	0.182
	10. Семантика	0.159	<0.01	0.943	0.00	-0.480	0.091	-0.224	0.182
Вар. 134	Total	0.034	<0.05	0.993	0.00	0.221	0.092	0.508	0.183
	1. Български	0.116	<0.01	0.972	0.00	-0.129	0.092	-0.629	0.183
	2. Литература	0.104	<0.01	0.972	0.00	0.078	0.092	-0.473	0.183
	3. История	0.126	<0.01	0.967	0.00	-0.139	0.092	-0.402	0.183
	4. География	0.117	<0.01	0.968	0.00	0.083	0.092	-0.112	0.183
	5. Математика	0.160	<0.01	0.944	0.00	0.581	0.092	-0.181	0.183
	6. Физика	0.173	<0.01	0.927	0.00	0.796	0.092	1.510	0.183
	7. Химия	0.156	<0.01	0.942	0.00	0.650	0.092	1.084	0.183
	8. Биология	0.143	<0.01	0.956	0.00	0.379	0.092	0.050	0.183
	9. Разсъждения	0.142	<0.01	0.944	0.00	-0.511	0.092	-0.559	0.183
	10. Семантика	0.160	<0.01	0.933	0.00	-0.691	0.092	0.630	0.183
Вар. 141	Total	0.042	<0.01	0.996	0.04	0.133	0.090	0.141	0.181
	1. Български	0.184	<0.01	0.934	0.00	-0.667	0.090	0.176	0.181
	2. Литература	0.135	<0.01	0.967	0.00	0.152	0.090	-0.105	0.181
	3. История	0.143	<0.01	0.957	0.00	0.368	0.090	-0.084	0.181
	4. География	0.137	<0.01	0.962	0.00	0.237	0.090	-0.166	0.181
	5. Математика	0.114	<0.01	0.970	0.00	0.032	0.090	-0.700	0.181
	6. Физика	0.172	<0.01	0.936	0.00	0.513	0.090	0.213	0.181
	7. Химия	0.191	<0.01	0.931	0.00	0.406	0.090	-0.118	0.181
	8. Биология	0.153	<0.01	0.955	0.00	0.322	0.090	0.200	0.181
	9. Разсъждения	0.117	<0.01	0.973	0.00	0.061	0.090	-0.508	0.181
	10. Семантика	0.140	<0.01	0.954	0.00	-0.443	0.090	-0.008	0.181
Вар. 154	Total	0.054	<0.01	0.993	0.00	0.206	0.088	-0.208	0.177
	1. Български	0.128	<0.01	0.967	0.00	-0.300	0.088	-0.102	0.177
	2. Литература	0.110	<0.01	0.974	0.00	0.063	0.088	-0.363	0.177

Вариант на ТОП	Раздел	Колмогоров-Смирнов *D*	Lillefors *p*	Shapiro-Wilk's *W*	Shapiro-Wilk's *p*	Асиметрия	Станд. грешка на асиметрията	Ексцес	Станд. грешка на ексцеса
1	2	3	4	5	6	7	8	9	10
	3. История	0.140	<0.01	0.954	0.00	0.403	0.088	0.240	0.177
	4. География	0.139	<0.01	0.961	0.00	0.258	0.088	-0.077	0.177
	5. Математика	0.156	<0.01	0.938	0.00	0.775	0.088	0.551	0.177
	6. Физика	0.159	<0.01	0.956	0.00	0.227	0.088	-0.217	0.177
	7. Химия	0.160	<0.01	0.947	0.00	0.454	0.088	0.215	0.177
	8. Биология	0.131	<0.01	0.967	0.00	0.117	0.088	-0.420	0.177
	9. Разсъждения	0.135	<0.01	0.955	0.00	-0.521	0.088	0.102	0.177
	10. Семантика	0.163	<0.01	0.945	0.00	-0.528	0.088	0.197	0.177
Вар. 166	Total	0.043	<0.01	0.997	**0.16**	0.112	0.085	0.054	0.169
	1. Български	0.143	<0.01	0.951	0.00	-0.514	0.085	-0.189	0.169
	2. Литература	0.120	<0.01	0.969	0.00	0.316	0.085	-0.244	0.169
	3. История	0.143	<0.01	0.964	0.00	0.293	0.085	-0.236	0.169
	4. География	0.130	<0.01	0.964	0.00	0.092	0.085	-0.099	0.169
	5. Математика	0.143	<0.01	0.962	0.00	0.303	0.085	0.208	0.169
	6. Физика	0.166	<0.01	0.946	0.00	0.378	0.085	-0.194	0.169
	7. Химия	0.190	<0.01	0.936	0.00	0.437	0.085	0.388	0.169
	8. Биология	0.133	<0.01	0.961	0.00	0.258	0.085	0.115	0.169
	9. Разсъждения	0.114	<0.01	0.970	0.00	-0.050	0.085	-0.538	0.169
	10. Семантика	0.134	<0.01	0.950	0.00	-0.506	0.085	0.374	0.169
Вар. 171	Total	0.059	<0.01	0.984	0.00	0.508	0.085	0.673	0.170
	1. Български	0.125	<0.01	0.966	0.00	-0.172	0.085	-0.471	0.170
	2. Литература	0.114	<0.01	0.973	0.00	0.080	0.085	-0.405	0.170
	3. История	0.116	<0.01	0.966	0.00	0.268	0.085	-0.108	0.170
	4. География	0.123	<0.01	0.958	0.00	0.335	0.085	-0.224	0.170
	5. Математика	0.123	<0.01	0.972	0.00	0.052	0.085	-0.374	0.170
	6. Физика	0.170	<0.01	0.930	0.00	0.717	0.085	0.828	0.170
	7. Химия	0.179	<0.01	0.927	0.00	0.697	0.085	0.584	0.170
	8. Биология	0.149	<0.01	0.942	0.00	0.618	0.085	0.953	0.170
	9. Разсъждения	0.113	<0.01	0.971	0.00	0.149	0.085	-0.381	0.170
	10. Семантика	0.145	<0.01	0.961	0.00	-0.299	0.085	-0.196	0.170
Вар. 175	Total	0.042	<0.01	0.996	0.02	0.078	0.080	-0.088	0.161
	1. Български	0.111	<0.01	0.971	0.00	0.078	0.080	-0.568	0.161
	2. Литература	0.141	<0.01	0.965	0.00	0.250	0.080	-0.170	0.161
	3. История	0.141	<0.01	0.968	0.00	-0.137	0.080	-0.200	0.161
	4. География	0.121	<0.01	0.969	0.00	0.154	0.080	-0.196	0.161
	5. Математика	0.128	<0.01	0.965	0.00	0.277	0.080	-0.508	0.161
	6. Физика	0.143	<0.01	0.957	0.00	0.319	0.080	-0.023	0.161

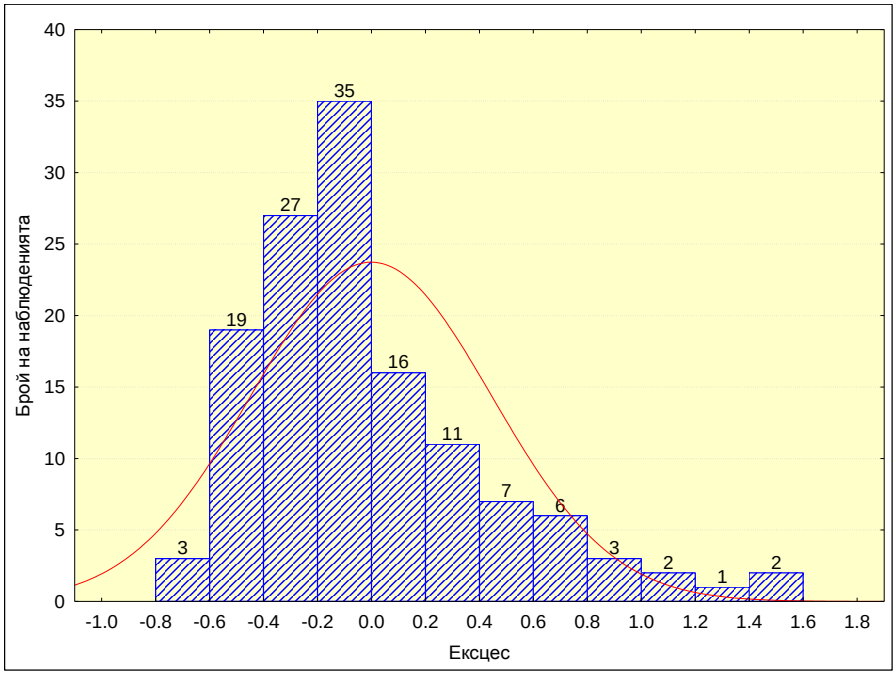
Вариант на ТОП	Раздел	Колмогоров-Смирнов *D*	Lillefors *p*	Shapiro-Wilk's *W*	Shapiro-Wilk's *p*	Асиметрия	Станд. грешка на асиметрията	Ексцес	Станд. грешка на ексцеса
1	2	3	4	5	6	7	8	9	10
	7. Химия	0.177	<0.01	0.945	0.00	0.571	0.080	0.666	0.161
	8. Биология	0.120	<0.01	0.967	0.00	0.003	0.080	-0.088	0.161
	9. Разсъждения	0.112	<0.01	0.970	0.00	0.191	0.080	-0.550	0.161
	10. Семантика	0.204	<0.01	0.907	0.00	-1.024	0.080	1.539	0.161
Вар. 192	Total	0.048	<0.01	0.992	0.00	0.309	0.077	-0.132	0.153
	1. Български	0.103	<0.01	0.975	0.00	0.066	0.077	-0.369	0.153
	2. Литература	0.143	<0.01	0.960	0.00	0.366	0.077	-0.322	0.153
	3. История	0.134	<0.01	0.965	0.00	0.258	0.077	-0.259	0.153
	4. География	0.146	<0.01	0.965	0.00	0.148	0.077	-0.213	0.153
	5. Математика	0.125	<0.01	0.972	0.00	0.210	0.077	-0.477	0.153
	6. Физика	0.148	<0.01	0.958	0.00	0.384	0.077	-0.059	0.153
	7. Химия	0.183	<0.01	0.926	0.00	0.600	0.077	0.484	0.153
	8. Биология	0.123	<0.01	0.965	0.00	0.252	0.077	-0.229	0.153
	9. Разсъждения	0.134	<0.01	0.967	0.00	0.304	0.077	-0.419	0.153
	10. Семантика	0.127	<0.01	0.970	0.00	-0.114	0.077	0.002	-0.114
Вар. 198	Total	0.048	<0.01	0.995	0.00	0.214	0.080	0.019	0.160
	1. Български	0.125	<0.01	0.971	0.00	-0.118	0.080	-0.441	0.160
	2. Литература	0.125	<0.01	0.969	0.00	0.091	0.080	-0.488	0.160
	3. История	0.112	<0.01	0.972	0.00	0.131	0.080	-0.436	0.160
	4. География	0.157	<0.01	0.950	0.00	0.436	0.080	0.032	0.160
	5. Математика	0.127	<0.01	0.970	0.00	0.328	0.080	-0.165	0.160
	6. Физика	0.157	<0.01	0.956	0.00	0.120	0.080	-0.493	0.160
	7. Химия	0.161	<0.01	0.942	0.00	0.548	0.080	0.765	0.160
	8. Биология	0.136	<0.01	0.959	0.00	0.164	0.080	-0.298	0.160
	9. Разсъждения	0.110	<0.01	0.973	0.00	-0.006	0.080	-0.375	0.160
	10. Семантика	0.171	<0.01	0.942	0.00	-0.413	0.080	0.203	0.160

ПРИЛОЖЕНИЕ 4.

Фигура 1. Хистограма на разпределението на индексите на асиметрия



Фигура 2. Хистограма на разпределението на индексите на ексцес



ПРИЛОЖЕНИЕ 5.

Матрица на Пиърсъновите коефициенти на корелация за данните от вар. 134, субтест 3. История

Въпрос	21	22	23	24	25	26	27	28	29	30
21	1.00	-0.06	0.05	0.03	0.05	0.14	0.12	0.02	0.16	0.08
22	-0.06	1.00	0.00	-0.02	-0.06	-0.03	-0.06	0.02	-0.09	-0.07
23	0.05	0.00	1.00	0.07	0.04	0.09	0.07	0.06	0.04	0.03
24	0.03	-0.02	0.07	1.00	0.11	0.07	0.06	0.04	-0.00	0.06
25	0.05	-0.06	0.04	0.11	1.00	0.17	0.12	0.03	0.24	0.10
26	0.14	-0.03	0.09	0.07	0.17	1.00	0.16	0.04	0.21	0.16
27	0.12	-0.06	0.07	0.06	0.12	0.16	1.00	0.07	0.14	0.07
28	0.02	0.02	0.06	0.04	0.03	0.04	0.07	1.00	0.00	-0.05
29	0.16	-0.09	0.04	-0.00	0.24	0.21	0.14	0.00	1.00	0.18
30	0.08	-0.07	0.03	0.06	0.10	0.16	0.07	-0.05	0.18	1.00
Means	0.78	0.04	0.29	0.32	0.79	0.63	0.63	0.23	0.80	0.57
Std.Dev.	0.41	0.20	0.45	0.47	0.41	0.48	0.48	0.42	0.40	0.50
No.Cases	713									
Matrix	1.00									

Матрица на тетрахоричните коефициенти на корелация за данните от вар. 134, субтест 3. История

Въпрос	21	22	23	24	25	26	27	28	29	30
21	1.00	-0.24	0.12	0.05	0.12	0.27	0.22	0.04	0.33	0.16
22	-0.24	1.00	0.02	-0.10	-0.26	-0.10	-0.21	0.08	-0.33	-0.28
23	0.12	0.02	1.00	0.13	0.09	0.16	0.12	0.12	0.09	0.05
24	0.05	-0.10	0.13	1.00	0.25	0.13	0.11	0.08	-0.00	0.10
25	0.12	-0.26	0.09	0.25	1.00	0.31	0.24	0.07	0.46	0.19
26	0.27	-0.10	0.16	0.13	0.31	1.00	0.26	0.08	0.40	0.25
27	0.22	-0.21	0.12	0.11	0.24	0.26	1.00	0.14	0.26	0.12
28	0.04	0.08	0.12	0.08	0.07	0.08	0.14	1.00	0.01	-0.09
29	0.33	-0.33	0.09	-0.00	0.46	0.40	0.26	0.01	1.00	0.33
30	0.16	-0.28	0.05	0.10	0.19	0.25	0.12	-0.09	0.33	1.00
Means	0.78	0.04	0.29	0.32	0.79	0.63	0.63	0.23	0.80	0.57
Std.Dev.	0.41	0.20	0.45	0.47	0.41	0.48	0.48	0.42	0.40	0.50
No.Cases	713									
Matrix	1.00									

ПРИЛОЖЕНИЕ 6.

Случай 1. Резултати от прилагането на анализа на главни фактори, метод на главните оси, върху данни от субтестовете на варианти 134, 141 и 171

(анализите са направени върху тетрахорични корелации, с използването на анализа на глави компоненти за определяне на симулираните собствени стойности по метода на Хорн)

Вариант 134	**Статистики на реалните данни**				**Статистики на симу-лираните данни**	
раздел	**пореден номер на фактор**	**собствена стойност**	**% от цялата дисперсия**	**кумул. % от цялата дисперсия**	**средни стойности**	**95-ти про-центил**
1	2	3	4	5	6	7
1. Български	**F1**	**2.464**	**24.637**	**24.637**	**1.190**	**1.242**
	F2	0.316	3.159	27.796	1.131	1.169
	F3	0.215	2.151	29.948	1.087	1.116
	F4	0.110	1.096	31.044	1.049	1.077
	F5	0.032	0.320	31.364	1.013	1.039
	F6				0.980	1.005
	F7				0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
2. Литература	**F1**	**1.389**	**13.891**	**13.891**	**1.190**	**1.242**
	F2	0.296	2.962	16.853	1.131	1.169
	F3	0.207	2.065	18.918	1.087	1.116
	F4	0.035	0.352	19.270	1.049	1.077
	F5	0.005	0.051	19.321	1.013	1.039
	F6				0.980	1.005
	F7				0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
3. История	**F1**	**1.903**	**19.034**	**19.034**	**1.190**	**1.242**
	F2	0.416	4.165	23.199	1.131	1.169
	F3	0.217	2.166	25.365	1.087	1.116
	F4	0.104	1.039	26.403	1.049	1.077
	F5	0.077	0.775	27.178	1.013	1.039
	F6				0.980	1.005
	F7				0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
4. География	**F1**	**1.267**	**14.078**	**14.078**	**1.173**	**1.226**
(9 въпроса)	F2	0.352	3.910	17.987	1.118	1.155
	F3	0.285	3.162	21.150	1.072	1.103

Вариант 134	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
	F4	0.078	0.863	22.012	1.033	1.062
	F5	0.039	0.436	22.448	0.996	1.022
	F6				0.962	0.989
	F7				0.925	0.951
	F8				0.885	0.917
	F9				0.836	0.873
5. Математика	**F1**	**2.498**	**24.979**	**24.979**	**1.190**	**1.242**
	F2	0.780	7.804	32.783	1.131	1.169
	F3	0.285	2.855	35.638	1.087	1.116
	F4	0.216	2.162	37.799	1.049	1.077
	F5	0.086	0.858	38.657	1.013	1.039
	F6				0.980	1.005
	F7				0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
6. Физика	**F1**	**1.267**	**12.670**	**12.670**	**1.190**	**1.242**
	F2	0.423	4.229	16.899	1.131	1.169
	F3	0.356	3.561	20.460	1.087	1.116
	F4	0.264	2.644	23.103	1.049	1.077
	F5	0.172	1.723	24.826	1.013	1.039
	F6	0.031	0.310	25.136	0.980	1.005
	F7	0.020	0.202	25.337	0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
7. Химия	F1	1.119	11.193	11.193	1.190	1.242
	F2	0.478	4.780	15.973	1.131	1.169
	F3	0.273	2.731	18.704	1.087	1.116
	F4	0.211	2.114	20.818	1.049	1.077
	F5				1.013	1.039
	F6				0.980	1.005
	F7				0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
8. Биология	F1	0.872	8.719	8.719	1.190	1.242
	F2	0.333	3.328	12.047	1.131	1.169
	F3	0.182	1.824	13.871	1.087	1.116

Вариант 134	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
	F4	0.151	1.507	15.379	1.049	1.077
	F5	0.125	1.247	16.626	1.013	1.039
	F6				0.980	1.005
	F7				0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
9. Разсъждения	**F1**	**2.889**	**28.887**	**28.887**	**1.190**	**1.242**
	F2	0.508	5.080	33.967	1.131	1.169
	F3	0.168	1.680	35.646	1.087	1.116
	F4	0.057	0.567	36.213	1.049	1.077
	F5	0.012	0.117	36.330	1.013	1.039
	F6				0.980	1.005
	F7				0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862
10. Семантика	**F1**	**2.327**	**23.273**	**23.273**	**1.190**	**1.242**
	F2	0.503	5.027	28.300	1.131	1.169
	F3	0.297	2.970	31.270	1.087	1.116
	F4	0.153	1.527	32.797	1.049	1.077
	F5	0.090	0.903	33.700	1.013	1.039
	F6	0.025	0.251	33.951	0.980	1.005
	F7	0.018	0.183	34.134	0.945	0.969
	F8				0.910	0.938
	F9				0.871	0.903
	F10				0.824	0.862

Вариант 141	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
1. Български	**F1**	**2.269**	**22.686**	**22.686**	**1.186**	**1.236**
	F2	0.712	7.116	29.802	1.130	1.166
	F3	0.541	5.406	35.207	1.088	1.119
	F4	0.193	1.931	37.139	1.049	1.075
	F5	0.137	1.367	38.506	1.014	1.039
	F6	0.072	0.72	39.221	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863
2. Литература	F1	0.861	8.612	8.612	1.186	1.236
	F2	0.272	2.716	11.329	1.130	1.166
	F3	0.180	1.803	13.132	1.088	1.119
	F4	0.142	1.420	14.552	1.049	1.075
	F5	0.104	1.036	15.588	1.014	1.039
	F6				0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863
3. История	F1	1.027	10.270	10.270	1.186	1.236
	F2	0.665	6.650	16.920	1.130	1.166
	F3	0.296	2.963	19.884	1.088	1.119
	F4	0.178	1.777	21.660	1.049	1.075
	F5	0.061	0.611	22.271	1.014	1.039
	F6	0.039	0.389	22.660	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863
4. География	F1	0.940	9.405	9.405	1.186	1.236
	F2	0.485	4.852	14.257	1.130	1.166
	F3	0.330	3.301	17.558	1.088	1.119
	F4	0.250	2.500	20.057	1.049	1.075
	F5	0.116	1.163	21.220	1.014	1.039
	F6	0.002	0.023	21.243	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863

Вариант 141	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
5. Математика	**F1**	**2.487**	**24.866**	**24.866**	**1.186**	**1.236**
	F2	0.714	7.140	32.005	1.130	1.166
	F3	0.530	5.302	37.308	1.088	1.119
	F4	0.234	2.341	39.649	1.049	1.075
	F5	0.079	0.790	40.439	1.014	1.039
	F6	0.044	0.436	40.875	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863
6. Физика	F1	0.842	8.416	8.416	1.186	1.236
	F2	0.462	4.619	13.035	1.130	1.166
	F3	0.348	3.478	16.513	1.088	1.119
	F4	0.261	2.609	19.123	1.049	1.075
	F5	0.138	1.382	20.504	1.014	1.039
	F6	0.042	0.417	20.921	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863
7. Химия	F1	0.556	5.562	5.562	1.186	1.236
	F2	0.336	3.360	8.922	1.130	1.166
	F3	0.289	2.888	11.811	1.088	1.119
	F4	0.247	2.465	14.276	1.049	1.075
	F5	0.172	1.718	15.994	1.014	1.039
	F6	0.028	0.282	16.276	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863
8. Биология	F1	0.827	8.272	8.272	1.186	1.236
	F2	0.354	3.535	11.807	1.130	1.166
	F3	0.298	2.982	14.790	1.088	1.119
	F4	0.180	1.802	16.592	1.049	1.075
	F5	0.116	1.157	17.749	1.014	1.039
	F6	0.025	0.255	18.004	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903

Вариант 141	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
	F10				0.826	0.863
9. Разсъждения	**F1**	**2.059**	**20.592**	**20.592**	**1.186**	**1.236**
	F2	0.471	4.713	25.305	1.130	1.166
	F3	0.303	3.031	28.336	1.088	1.119
	F4	0.136	1.360	29.696	1.049	1.075
	F5	0.088	0.88	30.576	1.014	1.039
	F6				0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863
10. Семантика	**F1**	**2.382**	**23.820**	**23.820**	**1.186**	**1.236**
	F2	0.708	7.078	30.898	1.130	1.166
	F3	0.425	4.255	35.153	1.088	1.119
	F4	0.347	3.468	38.621	1.049	1.075
	F5	0.139	1.387	40.008	1.014	1.039
	F6	0.087	0.87	40.879	0.979	1.002
	F7				0.946	0.970
	F8				0.911	0.937
	F9				0.872	0.903
	F10				0.826	0.863

Вариант 171	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
1. Български	**F1**	**1.862**	**18.625**	**18.625**	**1.174**	**1.220**
	F2	0.444	4.444	23.068	1.122	1.157
	F3	0.282	2.817	25.886	1.082	1.111
	F4	0.108	1.084	26.969	1.045	1.070
	F5	0.026	0.257	27.226	1.014	1.037
	F6	0.000	0.003	27.229	0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872
2. Литература	**F1**	**1.391**	**13.909**	**13.909**	**1.174**	**1.220**
	F2	0.367	3.671	17.580	1.122	1.157
	F3	0.290	2.896	20.476	1.082	1.111
	F4	0.128	1.279	21.755	1.045	1.070
	F5	0.088	0.879	22.634	1.014	1.037
	F6				0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872
3. История	**F1**	**1.492**	**14.922**	**14.922**	**1.174**	**1.220**
	F2	0.464	4.642	19.564	1.122	1.157
	F3	0.181	1.806	21.370	1.082	1.111
	F4	0.133	1.333	22.703	1.045	1.070
	F5	0.037	0.366	23.069	1.014	1.037
	F6				0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872
4. География	F1	1.079	10.792	10.792	1.174	1.220
	F2	0.628	6.281	17.073	1.122	1.157
	F3	0.173	1.734	18.807	1.082	1.111
	F4	0.140	1.399	20.207	1.045	1.070
	F5	0.039	0.387	20.594	1.014	1.037
	F6	0.030	0.300	20.894	0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872

Вариант 171	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
5. Математика	**F1**	**3.086**	**30.862**	**30.862**	**1.174**	**1.220**
	F2	0.715	7.146	38.008	1.122	1.157
	F3	0.373	3.732	41.740	1.082	1.111
	F4	0.265	2.646	44.386	1.045	1.070
	F5	0.103	1.027	45.413	1.014	1.037
	F6	0.032	0.321	45.734	0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872
6. Физика	F1	1.131	11.305	11.305	1.174	1.220
	F2	0.490	4.901	16.206	1.122	1.157
	F3	0.369	3.689	19.895	1.082	1.111
	F4	0.213	2.127	22.022	1.045	1.070
	F5	0.018	0.183	22.205	1.014	1.037
	F6				0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872
7. Химия	F1	1.014	10.142	10.142	1.174	1.220
	F2	0.426	4.264	14.406	1.122	1.157
	F3	0.371	3.707	18.113	1.082	1.111
	F4	0.151	1.506	19.619	1.045	1.070
	F5	0.086	0.864	20.483	1.014	1.037
	F6				0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872
8. Биология	**F1**	**1.266**	**12.657**	**12.657**	**1.174**	**1.220**
	F2	0.467	4.666	17.323	1.122	1.157
	F3	0.422	4.219	21.542	1.082	1.111
	F4	0.202	2.018	23.560	1.045	1.070
	F5	0.068	0.682	24.242	1.014	1.037
	F6	0.011	0.109	24.351	0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909

Вариант 171	Статистики на реалните данни				Статистики на симу-лираните данни	
раздел	пореден номер на фактор	собствена стойност	% от цялата дисперсия	кумул. % от цялата дисперсия	средни стойности	95-ти про-центил
1	2	3	4	5	6	7
	F10				0.836	0.872
9. Разсъждения	**F1**	**1.417**	**14.173**	**14.173**	**1.174**	**1.220**
	F2	0.493	4.927	19.101	1.122	1.157
	F3	0.251	2.511	21.612	1.082	1.111
	F4	0.155	1.552	23.164	1.045	1.070
	F5	0.068	0.683	23.847	1.014	1.037
	F6				0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872
10. Семантика	**F1**	**1.515**	**15.151**	**15.151**	**1.174**	**1.220**
	F2	0.416	4.165	19.316	1.122	1.157
	F3	0.275	2.746	22.062	1.082	1.111
	F4	0.108	1.081	23.142	1.045	1.070
	F5				1.014	1.037
	F6				0.982	1.005
	F7				0.949	0.973
	F8				0.916	0.941
	F9				0.880	0.909
	F10				0.836	0.872

ПРИЛОЖЕНИЕ 7.

Факторни матрици на въпросите в субтестовете от вариант 141

номер на въпрос	1. Български език	2. Литература	3. История	4. География	5. Математика	6. Физика	7. Химия	8. Биология	9. разсъждения	10. Семантика
1.	0.567	-0.195	0.476	0.449	0.573	0.107	-0.100	-0.123	0.340	0.307
2.	0.638	0.262	-0.035	0.254	0.484	0.003	-0.251	0.256	0.558	0.234
3.	0.176	0.052	0.105	-0.115	0.598	0.293	0.049	0.500	0.140	0.113
4.	0.452	0.183	0.464	0.273	0.330	-0.210	-0.240	0.227	0.663	0.445
5.	0.656	0.421	0.388	0.136	0.626	0.480	0.433	-0.168	0.386	0.390
6.	0.381	0.517	0.180	0.492	0.619	-0.193	-0.153	-0.045	0.331	0.612
7.	0.413	0.080	0.304	0.330	0.642	0.344	-0.409	0.532	0.384	0.534
8.	0.566	0.269	0.377	0.061	-0.005	0.088	0.046	0.174	0.125	0.592
9.	0.488	0.328	0.384	0.342	0.486	0.300	-0.142	0.311	0.561	0.769
10.	-0.021	0.297	0.086	0.312	0.184	0.465	0.151	0.060	0.661	0.518
обяснена дисперсия	2.269	0.861	1.027	0.940	2.487	0.842	0.556	0.827	2.059	2.382
% от цялата дисперсия	22.70	8.60	10.30	9.40	24.90	8.40	5.60	8.30	20.60	23.8

*Забележки: *Обяснена дисперсия – собствена стойност на съответния фактор*

*** Факторните тегла на въпросите във всички факторни решения, с изключени на тези по български език и математика, са умножени с коефициент -1.00.*

Факторни матрици на въпросите в субтестовете от вариант 171

номер на въпрос	1. Български език	2. Литература	3. История	4. География	5. Математика	6. Физика	7. Химия	8. Биология	9. разсъждения	10. Семантика
1.	0.621	0.158	-0.031	0.382	0.841	0.270	0.455	0.491	0.298	0.156
2.	0.434	0.232	0.238	0.039	0.418	0.433	0.284	0.576	-0.181	0.347
3.	0.094	0.313	0.514	0.228	0.577	0.432	0.262	0.313	0.377	0.232
4.	0.197	0.423	0.523	0.262	0.288	-0.237	-0.079	-0.077	0.063	0.538
5.	0.347	0.537	0.369	0.109	0.497	0.187	-0.074	0.396	0.364	0.603
6.	0.661	0.311	0.143	0.434	0.452	0.274	0.348	0.227	0.454	0.446
7.	0.588	0.488	0.505	-0.276	0.479	0.630	0.128	0.340	0.535	0.380
8.	-0.246	0.389	0.218	0.390	0.717	0.066	0.070	0.341	0.414	0.232
9.	0.077	-0.073	0.495	0.422	0.717	0.261	0.543	0.232	0.496	0.409
10.	0.521	0.506	0.438	0.453	0.281	0.220	0.457	0.309	0.328	0.314
обяснена дисперсия	1.862	1.391	1.492	1.079	3.086	1.131	1.014	1.266	1.417	1.515
% от цялата дисперсия	18.60	13.90	14.90	10.80	30.90	11.30	10.10	12.70	14.20	15.20

*Забележки: *Обяснена дисперсия – собствена стойност на съответния фактор*

*** Факторните тегла на въпросите във всички факторни решения, с изключени на тези по български език, литература и физика са умножени с коефициент -1.00.*

ПРИЛОЖЕНИЕ 8.

Случай 2. Резултати от прилагането на анализа на главни фактори, метод на главните оси, върху данни от варианти 134, 141 и 171

(анализите са направени върху тетрахорични корелации, с използването на анализа на глави компоненти за определяне на симулираните собствени стойности по метода на Хорн)

Собствени стойности на вариант 134

Вар. 134 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
1	12.278	12.658	12.278	12.658	1.826	1.880
2	2.851	2.939	15.129	15.596	1.775	1.817
3	2.406	2.480	17.534	18.077	1.734	1.771
4	2.229	2.298	19.764	20.375	1.700	1.731
5	2.055	2.119	21.819	22.494	1.670	1.700
6	1.934	1.994	23.754	24.488	1.640	1.668
7	1.908	1.967	25.662	26.455	1.615	1.640
8	1.714	1.767	27.376	28.223	1.590	1.616
9	1.619	1.669	28.995	29.892	1.567	1.590
10	**1.576**	**1.625**	**30.571**	**31.517**	**1.544**	**1.566**
11	1.482	1.528	32.053	33.044	1.522	1.545
12	1.406	1.449	33.459	34.494	1.502	1.524
13	1.355	1.397	34.814	35.890	1.482	1.504
14	1.259	1.298	36.073	37.188	1.462	1.484
15	1.232	1.271	37.305	38.459	1.443	1.462
16	1.207	1.244	38.512	39.703	1.424	1.442
17	1.153	1.189	39.665	40.892	1.406	1.424
18	1.112	1.146	40.777	42.038	1.388	1.406
19	1.044	1.076	41.820	43.114	1.371	1.388
20	0.989	1.020	42.809	44.133	1.353	1.371
21	0.955	0.984	43.764	45.118	1.337	1.354
22	0.911	0.939	44.675	46.057	1.320	1.338
23	0.859	0.885	45.533	46.942	1.305	1.321
24	0.788	0.813	46.322	47.754	1.289	1.306
25	0.773	0.797	47.094	48.551	1.273	1.290
26	0.754	0.777	47.848	49.328	1.258	1.274
27	0.727	0.750	48.576	50.078	1.243	1.259
28	0.673	0.694	49.249	50.772	1.229	1.244
29	0.635	0.655	49.884	51.427	1.214	1.228
30	0.626	0.646	50.510	52.072	1.199	1.215
31	0.586	0.604	51.096	52.676	1.185	1.199
32	0.575	0.592	51.670	53.268	1.171	1.185
33	0.548	0.565	52.219	53.834	1.158	1.173
34	0.503	0.518	52.721	54.352	1.144	1.158
35	0.476	0.491	53.198	54.843	1.131	1.145

Вар. 134 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
36	0.459	0.474	53.657	55.316	1.117	1.132
37	0.431	0.444	54.088	55.761	1.104	1.117
38	0.421	0.434	54.509	56.195	1.091	1.105
39	0.406	0.419	54.916	56.614	1.078	1.091
40	0.339	0.349	55.254	56.963	1.065	1.078
41	0.337	0.347	55.591	57.311	1.053	1.066
42	0.326	0.336	55.917	57.647	1.040	1.053
43	0.306	0.315	56.223	57.962	1.028	1.042
44	0.249	0.256	56.472	58.218	1.016	1.029
45	0.219	0.226	56.691	58.444	1.003	1.016
46	0.205	0.211	56.896	58.656	0.991	1.005
47	0.178	0.184	57.074	58.840	0.979	0.992
48	0.153	0.158	57.228	58.997	0.968	0.981
49	0.152	0.157	57.380	59.154	0.956	0.969
50	0.123	0.127	57.503	59.281	0.944	0.956
51	0.096	0.099	57.599	59.381	0.932	0.945
52	0.077	0.080	57.677	59.460	0.921	0.933
53	0.048	0.049	57.725	59.510	0.910	0.922
54	0.029	0.030	57.754	59.540	0.898	0.910
55	0.015	0.015	57.769	59.555	0.887	0.899
56					0.876	0.888
57					0.865	0.876
58					0.854	0.866
59					0.843	0.855
60					0.832	0.844
61					0.821	0.832
62					0.810	0.822
63					0.800	0.812
64					0.789	0.801
65					0.778	0.790
66					0.767	0.780
67					0.757	0.768
68					0.746	0.758
69					0.736	0.748
70					0.726	0.737
71					0.715	0.726
72					0.705	0.716
73					0.695	0.706
74					0.684	0.697
75					0.674	0.686
76					0.664	0.675
77					0.653	0.665
78					0.643	0.655
79					0.632	0.644

Вар. 134 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
80					0.622	0.634
81					0.612	0.623
82					0.601	0.613
83					0.590	0.602
84					0.579	0.591
85					0.569	0.581
86					0.558	0.570
87					0.547	0.559
88					0.536	0.548
89					0.525	0.538
90					0.514	0.527
91					0.503	0.516
92					0.491	0.503
93					0.478	0.491
94					0.464	0.478
95					0.450	0.465
96					0.434	0.451
97					0.414	0.433

Собствени стойности на вариант 141

Вар. 141 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
1	8.071	8.586	8.071	8.586	1.801	1.854
2	2.630	2.798	10.701	11.384	1.747	1.786
3	2.317	2.465	13.018	13.849	1.708	1.743
4	2.178	2.317	15.197	16.167	1.675	1.704
5	1.903	2.024	17.099	18.191	1.645	1.673
6	1.851	1.969	18.950	20.160	1.618	1.645
7	1.748	1.860	20.698	22.020	1.592	1.619
8	1.708	1.817	22.407	23.837	1.567	1.592
9	1.646	1.751	24.053	25.588	1.544	1.565
10	**1.567**	**1.667**	**25.620**	**27.255**	**1.522**	**1.546**
11	1.404	1.494	27.024	28.749	1.500	1.521
12	1.372	1.460	28.396	30.209	1.480	1.501
13	1.305	1.389	29.702	31.598	1.460	1.479
14	1.229	1.308	30.931	32.906	1.440	1.460
15	1.199	1.275	32.130	34.181	1.422	1.441
16	1.180	1.255	33.310	35.436	1.403	1.421
17	1.077	1.145	34.386	36.581	1.385	1.404
18	0.996	1.060	35.383	37.641	1.368	1.384
19	0.955	1.016	36.338	38.657	1.351	1.368
20	0.942	1.002	37.280	39.660	1.334	1.353
21	0.915	0.974	38.195	40.633	1.318	1.335
22	0.879	0.935	39.075	41.569	1.302	1.319
23	0.827	0.880	39.902	42.449	1.286	1.303
24	0.791	0.842	40.693	43.291	1.271	1.288
25	0.762	0.811	41.455	44.101	1.255	1.271
26	0.719	0.765	42.174	44.866	1.240	1.256
27	0.699	0.744	42.873	45.610	1.225	1.240
28	0.628	0.668	43.501	46.278	1.210	1.226
29	0.599	0.637	44.100	46.915	1.196	1.211
30	0.573	0.610	44.674	47.525	1.182	1.196
31	0.549	0.584	45.223	48.110	1.168	1.183
32	0.533	0.567	45.756	48.677	1.154	1.169
33	0.505	0.537	46.261	49.214	1.141	1.154
34	0.474	0.504	46.735	49.718	1.127	1.141
35	0.436	0.464	47.171	50.182	1.114	1.127
36	0.405	0.431	47.576	50.613	1.100	1.114
37	0.374	0.398	47.951	51.011	1.087	1.102
38	0.335	0.356	48.286	51.368	1.075	1.089
39	0.317	0.337	48.603	51.705	1.062	1.076
40	0.295	0.314	48.898	52.019	1.049	1.062
41	0.273	0.290	49.171	52.309	1.036	1.050
42	0.264	0.281	49.435	52.590	1.024	1.037

Вар. 141 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
43	0.239	0.254	49.674	52.844	1.012	1.026
44	0.191	0.203	49.865	53.048	1.000	1.013
45	0.182	0.194	50.047	53.242	0.988	1.000
46	0.164	0.174	50.211	53.416	0.976	0.988
47	0.147	0.157	50.358	53.572	0.964	0.977
48	0.122	0.130	50.480	53.702	0.952	0.965
49	0.097	0.104	50.577	53.806	0.940	0.954
50	0.046	0.049	50.624	53.855	0.929	0.941
51	0.000	0.000	50.624	53.855	0.917	0.930
52					0.906	0.918
53					0.895	0.907
54					0.884	0.896
55					0.872	0.884
56					0.861	0.873
57					0.850	0.862
58					0.839	0.851
59					0.828	0.840
60					0.817	0.830
61					0.806	0.819
62					0.796	0.808
63					0.785	0.797
64					0.774	0.786
65					0.763	0.775
66					0.753	0.765
67					0.742	0.755
68					0.732	0.745
69					0.721	0.733
70					0.711	0.723
71					0.701	0.712
72					0.690	0.702
73					0.680	0.691
74					0.669	0.681
75					0.659	0.671
76					0.648	0.661
77					0.638	0.650
78					0.627	0.639
79					0.616	0.628
80					0.606	0.618
81					0.595	0.608
82					0.585	0.597
83					0.574	0.585
84					0.562	0.575
85					0.551	0.563
86					0.540	0.552

Вар. 141 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
87					0.528	0.542
88					0.516	0.530
89					0.505	0.518
90					0.492	0.506
91					0.478	0.492
92					0.464	0.479
93					0.447	0.464
94					0.427	0.446

Собствени стойности на вариант 171

Вар. 171 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
1	8.601	8.688	8.601	8.688	1.767	1.818
2	2.466	2.491	11.067	11.178	1.720	1.760
3	2.348	2.372	13.415	13.551	1.683	1.714
4	2.017	2.037	15.432	15.588	1.653	1.682
5	1.887	1.906	17.319	17.494	1.625	1.652
6	1.871	1.890	19.191	19.384	1.599	1.624
7	1.683	1.700	20.873	21.084	1.575	1.598
8	**1.626**	**1.642**	**22.499**	**22.726**	**1.554**	**1.575**
9	1.546	1.561	24.045	24.287	1.532	1.554
10	1.517	1.532	25.561	25.819	1.511	1.533
11	1.413	1.427	26.974	27.246	1.491	1.511
12	1.342	1.355	28.315	28.601	1.473	1.491
13	1.265	1.277	29.580	29.879	1.454	1.474
14	1.194	1.207	30.774	31.085	1.436	1.454
15	1.143	1.155	31.918	32.240	1.419	1.438
16	1.128	1.140	33.046	33.380	1.402	1.420
17	1.088	1.099	34.134	34.479	1.385	1.403
18	1.058	1.069	35.192	35.548	1.369	1.386
19	1.040	1.050	36.232	36.598	1.354	1.369
20	0.937	0.947	37.169	37.545	1.338	1.354
21	0.891	0.900	38.060	38.444	1.323	1.339
22	0.882	0.891	38.942	39.336	1.308	1.324
23	0.834	0.842	39.776	40.178	1.294	1.309
24	0.829	0.837	40.605	41.015	1.279	1.294
25	0.748	0.755	41.353	41.770	1.265	1.279
26	0.722	0.729	42.074	42.499	1.251	1.265
27	0.698	0.705	42.772	43.204	1.237	1.251
28	0.643	0.650	43.415	43.854	1.224	1.238
29	0.632	0.638	44.047	44.492	1.210	1.224
30	0.614	0.621	44.661	45.112	1.197	1.211
31	0.581	0.587	45.242	45.699	1.184	1.197
32	0.557	0.563	45.799	46.262	1.171	1.184
33	0.532	0.537	46.331	46.799	1.158	1.171
34	0.490	0.495	46.821	47.294	1.145	1.158
35	0.471	0.476	47.292	47.770	1.133	1.146
36	0.438	0.442	47.730	48.212	1.121	1.134
37	0.408	0.413	48.138	48.624	1.109	1.122
38	0.377	0.381	48.515	49.005	1.097	1.109
39	0.355	0.358	48.870	49.363	1.085	1.098
40	0.307	0.310	49.177	49.674	1.073	1.085
41	0.295	0.298	49.472	49.972	1.062	1.074
42	0.289	0.292	49.761	50.264	1.050	1.062

Вар. 171 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
43	0.268	0.271	50.029	50.535	1.038	1.051
44	0.217	0.219	50.246	50.754	1.027	1.039
45	0.201	0.203	50.447	50.957	1.016	1.028
46	0.193	0.195	50.640	51.152	1.005	1.017
47	0.185	0.187	50.825	51.338	0.994	1.006
48	0.153	0.155	50.978	51.493	0.983	0.995
49	0.141	0.142	51.119	51.635	0.972	0.983
50	0.101	0.102	51.220	51.737	0.961	0.974
51	0.083	0.084	51.303	51.821	0.950	0.963
52	0.071	0.072	51.374	51.893	0.940	0.951
53	0.063	0.064	51.437	51.957	0.929	0.941
54	0.036	0.037	51.474	51.994	0.919	0.930
55	0.010	0.011	51.484	52.004	0.909	0.920
56					0.898	0.908
57					0.888	0.899
58					0.878	0.888
59					0.867	0.878
60					0.857	0.868
61					0.847	0.858
62					0.837	0.848
63					0.827	0.838
64					0.817	0.828
65					0.807	0.819
66					0.797	0.809
67					0.787	0.798
68					0.778	0.789
69					0.768	0.779
70					0.758	0.769
71					0.748	0.759
72					0.739	0.749
73					0.729	0.740
74					0.719	0.730
75					0.710	0.720
76					0.700	0.711
77					0.690	0.702
78					0.681	0.692
79					0.671	0.683
80					0.661	0.673
81					0.651	0.663
82					0.642	0.652
83					0.632	0.643
84					0.622	0.633
85					0.612	0.623
86					0.602	0.613

Вар. 171 общ бал	Статистики на реалните данни				Статистики на симу-лираните данни	
пореден номер на фактор	собствени стойности	% от цялата дисперсия	кумулат. собствени стойности	кумулат. % от цялата дисперсия	средни стойности	95-ти процентил
87					0.592	0.604
88					0.582	0.594
89					0.572	0.583
90					0.561	0.573
91					0.551	0.562
92					0.540	0.551
93					0.529	0.541
94					0.517	0.530
95					0.505	0.519
96					0.492	0.506
97					0.479	0.493
98					0.463	0.479
99					0.443	0.461

ПРИЛОЖЕНИЕ 9.

Коефициенти на корелация r_{tet} > ±0.50

Вариант 134

No.	Двойки въпроси	r_{tet}
1.	33 (раздел 4. География) и 34 (раздел 4. География)	1.000
2.	42 (раздел 5. Математика) и 45 (раздел 5. Математика)	0.573
3.	81 (раздел 9. Разсъждения) и 84 (раздел 9. Разсъждения)	0.545
4.	33 (раздел 4. География), 77 (раздел 8. Биология)	0.533
5.	34 (раздел 4. География) и 77 (раздел 8. Биология)	0.532
6.	32 (раздел 4. География) и 41 (раздел 5. Математика)	0.533
7.	58 (раздел 6. Физика) и 81 (раздел 9. Разсъждения)	0.529
8.	21 (раздел 3. История) и 98 (раздел 10. Семантика)	0.514
9.	97 (раздел 10. Семантика) и 98 (раздел 10. Семантика)	0.513
10.	17 (раздел 2. Литература) и 41 (раздел 5. Математика)	0.504
11.	22 (раздел 2. История) и 81 (раздел 9. Разсъждения)	-0.507
12.	22 (раздел 2. История) и 38 (раздел 4. География)	-0.612

Вариант 141

No.	Двойки въпроси	r_{tet}
1.	1 (раздел 1. Литература) и 26 (раздел 3. История)	0.607
2.	98 (раздел 10. Семантика) и 99 (раздел 10. Семантика)	0.561
3.	48 (раздел 5. Математика) и 75 (раздел 8. Биология)	0.560
4.	1 (раздел 1. Литература) и 99 (раздел 10. Семантика)	0.545
5.	84 (раздел 9. Разсъждения) и 90 (раздел 9. Разсъждения)	0.538
6.	46 (раздел 5. Математика) и 47 (раздел 5. Математика)	0.533
7.	27 (раздел 3. История) и 98 (раздел 10. Семантика))	0.500
8.	48 (раздел 5. Математика) и 59 (раздел 6. Физика)	-1.000
9.	48 (раздел 5. Математика) и 54 (раздел 6. Физика)	-0.659
10.	38 (раздел 4. География) и 48 (раздел 5. Математика)	-0.653

Вариант 171

No.	Двойки въпроси	r_{tet}
1.	41 (раздел 5. Математика) и 48 (раздел 5. Математика)	0.686
2.	41 (раздел 5. Математика) и 49 (раздел 5. Математика)	0.652
3.	48 (раздел 5. Математика) и 49 (раздел 5. Математика)	0.646
4.	41 (раздел 5. Математика) и 70 (раздел 7. Химия)	0.643
5.	41 (раздел 5. Математика) и 43 (раздел 5. Математика)	0.571
6.	41 (раздел 5. Математика) и 63 (раздел 7. Химия)	0.560
7.	41 (раздел 5. Математика) и 95 (раздел 10. Семантика)	0.527
8.	41 (раздел 5. Математика) и 82 (раздел 9. Разсъждения)	-0.500

ПРИЛОЖЕНИЕ 10.

Матрица на факторните тегла на въпросите от вариант 134
Метод на ротация: Varimax normalized

Въпроси	Първични фактори									
	1	2	3	4	5	6	7	8	9	10
1	0.249	0.024	0.387	0.145	0.256	-0.068	0.073	-0.104	0.086	0.069
2	0.249	0.183	0.418	0.054	0.012	-0.024	0.048	0.029	0.184	-0.086
3	0.429	0.030	0.376	-0.115	0.076	0.152	0.274	0.099	0.163	0.111
4	0.206	0.088	0.534	-0.114	0.080	0.032	-0.010	0.157	0.110	0.113
5	0.237	0.024	0.251	0.024	0.037	-0.011	0.119	0.033	0.094	-0.030
6	0.349	0.171	0.464	-0.297	0.158	0.343	-0.041	0.073	-0.046	-0.088
7	0.111	0.114	0.396	0.092	-0.075	-0.085	-0.024	0.070	0.174	-0.217
8	0.333	-0.054	0.389	-0.100	0.159	0.089	0.063	0.075	0.113	0.045
9	0.254	-0.059	0.249	0.087	0.031	-0.001	-0.102	-0.025	0.188	-0.055
10	0.035	-0.100	0.328	0.152	0.170	0.057	0.092	0.143	0.045	0.011
11	0.057	0.124	0.153	-0.025	-0.119	-0.018	0.036	0.116	0.269	0.057
12	0.097	-0.016	0.133	0.033	0.026	-0.067	0.073	0.043	0.350	0.156
13	0.177	0.110	0.248	-0.051	-0.042	0.213	-0.034	0.035	0.236	0.056
14	-0.056	0.009	-0.194	-0.042	0.084	0.113	-0.210	-0.124	-0.066	-0.152
15	0.074	-0.009	0.425	0.007	0.094	0.146	0.001	-0.055	0.184	0.005
16	-0.023	0.025	0.207	-0.027	-0.053	-0.028	0.077	-0.036	0.296	-0.024
17	0.198	-0.122	0.222	-0.038	0.036	0.237	0.031	-0.096	0.167	0.114
18	0.143	-0.105	0.187	0.006	0.058	0.237	-0.006	0.190	0.289	0.119
19	0.216	0.008	0.118	0.075	0.030	0.312	0.086	-0.094	0.105	0.132
20	0.183	0.025	0.041	-0.206	-0.071	0.213	0.196	-0.004	0.272	-0.037
21	0.176	0.060	0.189	0.111	0.179	0.331	0.086	-0.017	0.105	0.138
22	-0.126	-0.081	-0.129	-0.885	-0.343	-0.169	-0.048	0.189	-0.215	-0.061
23	0.051	0.050	0.137	-0.102	0.101	-0.004	0.094	0.065	0.173	0.046
24	0.036	0.001	0.082	-0.046	0.066	0.108	0.259	-0.060	0.056	0.125
25	0.173	-0.028	0.179	0.056	0.083	0.109	0.158	-0.169	0.433	-0.048
26	0.210	-0.065	0.167	-0.083	0.111	0.154	0.141	0.052	0.475	0.172
27	-0.046	0.072	0.063	0.080	-0.031	0.149	0.059	0.032	0.432	0.046
28	-0.052	-0.109	0.276	0.068	-0.094	-0.034	0.082	0.197	0.094	0.358
29	0.248	0.041	0.326	0.070	0.289	0.234	0.106	-0.222	0.389	-0.125
30	0.038	0.154	0.191	0.034	0.221	-0.041	0.135	-0.123	0.396	0.076
31	0.185	0.198	0.000	0.071	0.017	0.053	-0.078	-0.029	0.333	-0.059
32	0.198	0.260	0.112	0.181	-0.044	0.111	0.095	0.027	0.214	0.051
34	0.154	-0.139	0.069	0.016	0.312	0.000	0.041	0.707	0.170	0.054
35	0.166	-0.030	0.056	0.062	0.087	0.058	0.031	0.144	0.333	-0.041
36	0.068	0.276	-0.054	0.024	0.020	0.032	-0.025	-0.042	0.107	-0.016
37	0.032	0.005	-0.130	0.133	0.028	0.443	-0.055	-0.040	0.024	0.081
38	0.042	0.130	0.131	0.416	-0.048	0.097	0.102	0.068	0.289	-0.091
39	0.162	0.048	0.198	0.154	0.050	0.276	-0.066	0.003	0.259	0.267
40	0.090	0.013	0.156	0.050	0.109	0.452	0.035	0.105	0.438	0.011
41	0.457	0.067	0.031	0.080	0.297	0.661	0.214	-0.159	0.038	-0.135
43	0.365	0.001	0.031	0.204	-0.035	-0.004	0.245	0.145	0.163	-0.202
44	0.406	-0.011	0.080	0.027	0.269	-0.031	0.297	0.006	0.047	-0.178

Въпроси	Първични фактори									
	1	2	3	4	5	6	7	8	9	10
45	0.592	0.188	0.096	0.192	-0.019	-0.001	0.239	0.035	-0.004	0.143
46	0.458	-0.107	0.111	0.017	0.039	0.073	0.245	-0.054	0.044	0.014
47	0.209	0.025	-0.080	0.152	0.030	-0.386	-0.073	-0.060	-0.042	0.299
48	0.411	0.024	0.039	0.006	0.138	0.086	-0.042	0.107	0.109	-0.110
49	0.430	0.197	-0.017	0.161	-0.226	-0.074	0.186	0.093	0.065	0.085
50	0.285	-0.132	0.111	0.052	0.367	0.050	-0.011	0.027	0.100	-0.138
51	0.114	0.061	-0.003	-0.101	-0.046	0.148	-0.164	-0.006	-0.035	0.163
52	0.167	0.334	-0.245	-0.139	0.223	0.048	0.300	0.051	0.166	-0.018
53	0.197	0.289	0.292	-0.204	-0.225	0.006	0.251	0.173	0.048	0.114
54	0.062	-0.168	-0.077	-0.063	-0.055	0.078	-0.017	0.072	0.043	0.264
55	0.081	0.513	0.008	-0.278	0.076	-0.160	0.170	0.003	0.113	0.053
56	-0.156	-0.017	0.194	-0.206	0.165	-0.042	0.247	-0.327	-0.169	-0.131
57	0.235	0.145	0.140	0.109	0.084	0.150	0.240	0.025	-0.054	0.031
58	0.237	0.147	-0.013	0.001	0.001	-0.053	0.456	0.017	0.136	0.149
59	0.427	-0.129	0.161	0.177	-0.094	0.040	0.329	0.027	0.076	-0.199
60	0.428	-0.177	0.096	-0.065	-0.036	0.023	0.306	-0.232	-0.053	0.075
61	0.158	0.168	0.013	-0.005	-0.038	0.319	0.395	0.255	0.253	-0.075
62	0.103	0.466	0.075	0.202	-0.184	0.068	0.107	-0.067	0.051	0.047
63	0.051	0.161	0.018	0.185	0.043	0.076	0.605	0.073	0.175	-0.172
64	0.115	-0.155	-0.014	-0.023	-0.038	-0.106	0.080	-0.098	0.034	-0.076
65	0.042	-0.019	0.116	0.162	-0.080	0.054	-0.092	-0.136	0.035	0.048
66	0.122	0.244	0.100	0.005	-0.173	0.246	0.225	0.104	-0.121	-0.037
67	0.064	-0.004	-0.089	0.083	0.154	0.010	-0.081	-0.369	0.118	0.054
68	0.012	0.565	0.093	0.063	0.128	0.096	0.042	0.017	-0.122	-0.171
69	-0.021	0.035	0.190	0.055	-0.007	0.122	0.028	0.009	0.124	-0.052
70	0.054	0.030	0.215	0.082	-0.024	0.092	-0.018	-0.173	0.021	-0.028
71	0.096	0.011	0.213	-0.121	-0.064	0.022	0.387	0.198	0.105	-0.068
72	0.041	-0.059	0.016	0.031	0.231	0.065	-0.005	0.051	0.043	0.024
73	-0.081	-0.094	0.216	0.035	0.165	0.080	0.146	-0.080	0.021	-0.228
74	-0.084	-0.007	-0.028	0.007	-0.152	0.068	0.116	0.152	0.174	0.047
75	0.052	-0.076	-0.105	0.095	0.138	-0.111	-0.077	-0.010	0.343	-0.054
76	0.034	-0.056	0.374	0.060	0.004	0.016	0.125	0.014	-0.002	0.034
77	0.196	0.065	0.158	0.275	0.256	0.298	0.076	0.347	0.080	0.113
78	0.068	0.128	0.119	0.281	0.018	0.279	0.101	0.199	-0.091	0.004
79	0.294	0.177	0.016	-0.048	-0.010	-0.071	0.177	0.321	0.184	-0.174
80	0.182	0.028	0.169	0.057	-0.045	0.202	0.150	0.165	0.006	-0.142
82	0.494	0.096	0.250	-0.048	0.087	0.167	0.086	-0.014	0.068	0.123
83	0.429	0.124	0.141	0.117	-0.007	0.093	-0.076	-0.020	0.029	0.146
84	0.628	0.138	0.060	0.001	0.079	0.198	-0.026	0.158	0.061	0.052
85	0.352	-0.004	0.127	0.022	0.052	0.053	0.030	-0.061	0.283	-0.143
86	0.247	0.102	0.181	0.068	-0.033	0.214	0.140	-0.142	0.265	0.076
87	0.445	-0.057	0.015	-0.057	0.082	0.152	0.073	-0.040	0.280	0.054
88	0.308	0.101	0.194	0.025	0.206	0.079	0.004	-0.079	0.139	-0.084
89	0.455	0.031	0.315	0.038	0.081	-0.043	-0.105	-0.060	0.167	0.136
90	0.238	-0.006	0.345	0.068	0.106	0.214	0.102	-0.092	0.217	0.061
91	0.059	0.148	0.380	-0.017	0.422	-0.051	0.262	-0.094	0.240	0.386
92	0.468	0.058	0.051	-0.078	0.053	-0.002	-0.007	0.132	0.175	0.140

Въпроси	Първични фактори									
	1	2	3	4	5	6	7	8	9	10
93	0.395	0.106	-0.033	0.005	-0.035	0.195	0.001	-0.142	0.051	0.063
94	0.271	-0.230	0.111	0.106	0.233	0.216	0.299	0.029	-0.089	0.144
95	0.323	-0.011	-0.229	-0.152	0.311	0.266	0.062	-0.030	0.115	0.113
96	0.398	0.048	0.087	-0.043	0.171	0.005	0.074	0.089	-0.074	0.125
97	0.232	0.187	0.112	-0.018	0.614	0.449	0.002	0.094	0.125	0.236
98	0.160	0.013	0.013	0.085	0.189	0.182	0.075	-0.176	0.066	0.851
99	0.332	0.002	0.190	0.016	-0.081	0.341	0.000	0.030	0.204	-0.008
100	0.025	0.236	0.107	-0.009	0.690	-0.087	0.019	-0.058	-0.024	-0.041
Expl.Var	6.142	2.133	3.858	2.112	2.840	3.147	2.644	1.982	3.446	2.266
Prp.Totl	0.063	0.022	0.040	0.022	0.029	0.032	0.027	0.020	0.036	0.023

ПРИЛОЖЕНИЕ 11.

Йерархичен факторен анализ.
Корелации между клъстърите от въпроси и факторите от първи и втори ред върху данни от вариант 134

Фактори	Клъстъри									
	1	2	3	4	5	6	7	8	9	10
S1	0.154	0.214	0.223	0.496	0.386	0.538	0.106	0.029	0.239	0.049
S2	0.254	-0.070	0.093	0.104	0.049	0.154	0.104	0.416	0.225	0.448
S3	0.357	0.518	0.237	0.322	0.116	0.228	0.657	0.100	0.399	-0.068
S4	0.638	0.180	0.729	0.273	0.487	0.523	0.259	0.050	0.598	0.092
S5	-0.072	-0.086	0.037	0.021	-0.186	0.050	0.066	-0.028	0.153	0.037
P1	0.610	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P2	0.000	0.801	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P3	0.000	0.000	0.594	0.000	0.000	0.000	0.000	0.000	0.000	0.000
P4	0.000	0.000	0.000	0.751	0.000	0.000	0.000	0.000	0.000	0.000
P5	0.000	0.000	0.000	0.000	0.750	0.000	0.000	0.000	0.000	0.000
P6	0.000	0.000	0.000	0.000	0.000	0.599	0.000	0.000	0.000	0.000
P7	0.000	0.000	0.000	0.000	0.000	0.000	0.689	0.000	0.000	0.000
P8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.902	0.000	0.000
P9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.594	0.000
P10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.884

Забележка: S – фактор от втори ред (secondary); P – фактор от първи ред (primary)

ПРИЛОЖЕНИЕ 12.

Йерархичен факторен анализ. Факторни тегла на въпросите по факторите от втори ред върху данни от вариант 134

Номер на въпрос в теста	Пореден номер на въпрос	Фактори от втори ред				
		1	2	3	4	5
1	1	0.240	0.139	0.264	0.160	-0.054
2	2	0.294	0.151	0.177	0.168	-0.002
3	3	0.408	0.043	0.257	0.247	-0.004
4	4	0.302	0.036	0.283	0.094	-0.015
5	5	0.225	0.069	0.125	0.122	-0.008
6	6	0.251	0.153	0.328	0.088	-0.085
7	7	0.206	0.146	0.102	0.074	0.028
8	8	0.294	0.048	0.292	0.104	-0.030
9	9	0.169	0.065	0.210	0.044	0.009
10	10	0.127	0.049	0.227	0.150	-0.004
11	11	0.184	-0.014	0.050	0.112	0.068
12	12	0.210	-0.034	0.127	0.131	0.066
13	13	0.190	0.049	0.218	0.143	0.045
14	14	-0.187	0.058	-0.011	-0.109	-0.042
15	15	0.176	0.128	0.274	0.097	0.037
16	16	0.160	0.065	0.056	0.061	0.088
17	17	0.151	0.031	0.255	0.084	0.047
18	18	0.168	-0.053	0.291	0.167	0.059
19	19	0.117	0.055	0.226	0.209	0.025
20	20	0.208	0.037	0.059	0.126	0.070
21	21	0.129	0.084	0.303	0.281	-0.009
22	22	-0.021	-0.246	-0.382	-0.549	0.018
23	23	0.161	0.018	0.086	0.087	0.009
24	24	0.119	0.031	0.053	0.143	0.024
25	25	0.239	0.147	0.189	0.198	0.083
26	26	0.297	-0.017	0.264	0.215	0.081
27	27	0.120	0.036	0.123	0.212	0.114
28	28	0.135	-0.161	0.134	0.082	0.079
29	29	0.262	0.274	0.349	0.261	0.018
30	30	0.228	0.138	0.157	0.219	0.031
31	31	0.138	0.086	0.107	0.167	0.019
32	32	0.188	0.085	0.115	0.306	0.021
34	33	0.195	-0.201	0.248	0.215	-0.064
35	34	0.174	0.006	0.163	0.171	0.033
36	35	0.048	0.078	-0.002	0.129	-0.022
37	36	-0.111	0.016	0.173	0.175	0.020
38	37	0.114	0.124	0.124	0.344	0.072
39	38	0.135	-0.000	0.334	0.245	0.048
40	39	0.155	0.074	0.344	0.309	0.088
41	40	0.137	0.248	0.356	0.404	-0.070
43	41	0.241	0.065	0.038	0.260	0.002

Номер на въпрос в теста	Пореден номер на въпрос	Фактори от втори ред				
		1	2	3	4	5
44	42	0.266	0.136	0.097	0.206	-0.092
45	43	0.326	0.023	0.113	0.332	-0.071
46	44	0.260	0.036	0.130	0.155	-0.022
47	45	0.070	-0.128	-0.035	-0.024	-0.066
48	46	0.179	0.039	0.183	0.128	-0.071
49	47	0.253	-0.037	-0.041	0.243	-0.005
50	48	0.134	0.095	0.256	0.097	-0.086
51	49	-0.008	-0.064	0.097	-0.026	-0.016
52	50	0.173	0.073	-0.069	0.289	-0.069
53	51	0.321	-0.009	-0.025	0.157	0.026
54	52	0.010	-0.178	0.069	-0.021	0.037
55	53	0.225	0.083	-0.132	0.137	-0.064
56	54	0.012	0.206	-0.079	-0.105	-0.025
57	55	0.169	0.095	0.111	0.273	-0.047
58	56	0.286	0.007	-0.057	0.292	0.009
59	57	0.274	0.092	0.063	0.209	0.021
60	58	0.230	0.030	0.043	0.044	-0.005
61	59	0.242	0.051	0.059	0.406	0.057
62	60	0.114	0.130	-0.038	0.291	0.010
63	61	0.229	0.169	-0.082	0.434	0.042
64	62	0.059	0.006	-0.047	-0.072	0.017
65	63	-0.007	0.043	0.095	0.020	0.035
66	64	0.105	0.067	-0.020	0.222	0.001
67	65	-0.038	0.105	0.077	0.001	-0.010
68	66	0.039	0.239	-0.014	0.248	-0.113
69	67	0.068	0.082	0.110	0.099	0.043
70	68	0.045	0.123	0.109	0.038	0.021
71	69	0.268	0.021	-0.018	0.165	0.046
72	70	0.019	0.011	0.142	0.073	-0.040
73	71	0.035	0.179	0.083	0.046	0.007
74	72	0.048	-0.071	-0.034	0.097	0.091
75	73	0.054	0.005	0.071	0.045	0.026
76	74	0.146	0.061	0.134	0.071	0.026
77	75	0.138	-0.001	0.348	0.410	-0.053
78	76	0.025	0.053	0.153	0.308	-0.021
79	77	0.273	0.006	-0.020	0.205	-0.029
80	78	0.140	0.066	0.103	0.183	0.004
82	79	0.301	0.054	0.262	0.199	-0.053
83	80	0.180	0.012	0.210	0.161	-0.050
84	81	0.269	-0.008	0.256	0.237	-0.097
85	82	0.233	0.110	0.164	0.123	0.010
86	83	0.222	0.111	0.186	0.244	0.059
87	84	0.250	0.008	0.208	0.158	0.004
88	85	0.196	0.154	0.217	0.151	-0.058
89	86	0.286	0.034	0.284	0.086	-0.038
90	87	0.236	0.125	0.303	0.212	0.031

Номер на въпрос в теста	Пореден номер на въпрос	Фактори от втори ред				
		1	2	3	4	5
91	88	0.312	0.088	0.282	0.284	-0.029
92	89	0.274	-0.075	0.170	0.130	-0.046
93	90	0.126	0.045	0.120	0.132	-0.027
94	91	0.142	-0.003	0.230	0.219	-0.042
95	92	0.089	-0.016	0.184	0.158	-0.077
96	93	0.201	-0.025	0.151	0.124	-0.107
97	94	0.138	0.090	0.488	0.387	-0.134
98	95	0.100	-0.181	0.304	0.242	-0.002
99	96	0.191	0.049	0.259	0.182	0.046
100	97	0.070	0.209	0.192	0.160	-0.196

ПРИЛОЖЕНИЕ 13.

Проверка на хипотезата за едномерност на тестов вариант 134.
Оценки на модела

Връзка	Оценки на модела			
	оценка на параметъра	стандартна грешка	Т-статистика	ниво на значимост
(Общ бал)-1->[въпрос 1]	0.423	0.032	13.241	0.000
(общ бал)-2->[въпрос 2]	0.423	0.032	13.223	0.000
(общ бал)-3->[въпрос 3]	0.638	0.023	27.175	0.000
(общ бал)-4->[въпрос 4]	0.432	0.032	13.641	0.000
(общ бал)-5->[въпрос 5]	0.326	0.035	9.394	0.000
(общ бал)-6->[въпрос 6]	0.531	0.028	18.878	0.000
(общ бал)-7->[въпрос 7]	0.240	0.037	6.558	0.000
(общ бал)-8->[въпрос 8]	0.485	0.030	16.208	0.000
(общ бал)-9->[въпрос 9]	0.302	0.035	8.570	0.000
(общ бал)-10->[въпрос 10]	0.269	0.036	7.472	0.000
(общ бал)-11->[въпрос 11]	0.198	0.037	5.326	0.000
(общ бал)-12->[въпрос 12]	0.259	0.036	7.176	0.000
(общ бал)-13->[въпрос 13]	0.371	0.034	11.078	0.000
(общ бал)-14->[въпрос 14]	-0.155	0.038	-4.095	0.000
(общ бал)-15->[въпрос 15]	0.356	0.034	10.480	0.000
(общ бал)-16->[въпрос 16]	0.174	0.038	4.641	0.000
(общ бал)-17->[въпрос 17]	0.356	0.034	10.503	0.000
(общ бал)-18->[въпрос 18]	0.371	0.034	11.079	0.000
(общ бал)-19->[въпрос 19]	0.373	0.033	11.152	0.000
(общ бал)-20->[въпрос 20]	0.318	0.035	9.113	0.000
(общ бал)-21->[въпрос 21]	0.438	0.032	13.896	0.000
(общ бал)-22->[въпрос 22]	-0.451	0.031	-14.516	0.000
(общ бал)-23->[въпрос 23]	0.207	0.037	5.583	0.000
(общ бал)-24->[въпрос 24]	0.207	0.037	5.587	0.000
(общ бал)-25->[въпрос 25]	0.428	0.032	13.457	0.000
(общ бал)-26->[въпрос 26]	0.496	0.030	16.796	0.000
(общ бал)-27->[въпрос 27]	0.244	0.036	6.680	0.000
(общ бал)-28->[въпрос 28]	0.134	0.038	3.518	0.000
(общ бал)-29->[въпрос 29]	0.603	0.025	24.018	0.000
(общ бал)-30->[въпрос 30]	0.348	0.034	10.196	0.000
(общ бал)-31->[въпрос 31]	0.269	0.036	7.484	0.000
(общ бал)-32->[въпрос 32]	0.356	0.034	10.507	0.000
(общ бал)-33->[въпрос 34]	0.296	0.035	8.358	0.000
(общ бал)-34->[въпрос 35]	0.303	0.035	8.593	0.000
(общ бал)-35->[въпрос 36]	0.107	0.038	2.795	0.005
(общ бал)-36->[въпрос 37]	0.141	0.038	3.706	0.000
(общ бал)-37->[въпрос 38]	0.279	0.036	7.792	0.000
(общ бал)-38->[въпрос 39]	0.418	0.032	13.002	0.000
(общ бал)-39->[въпрос 40]	0.483	0.030	16.124	0.000
(общ бал)-40->[въпрос 41]	0.665	0.022	29.995	0.000
(общ бал)-41->[въпрос 43]	0.360	0.034	10.624	0.000

Връзка	Оценки на модела			
	оценка на параметъра	стандартна грешка	Т-статистика	ниво на зна-чимост
(общ бал)-42->[въпрос 44]	0.426	0.032	13.369	0.000
(общ бал)-43->[въпрос 45]	0.528	0.028	18.664	0.000
(общ бал)-44->[въпрос 46]	0.430	0.032	13.556	0.000
(общ бал)-45->[въпрос 47]	-0.025	0.039	-0.656	0.512
(общ бал)-46->[въпрос 48]	0.368	0.034	10.946	0.000
(общ бал)-47->[въпрос 49]	0.303	0.035	8.587	0.000
(общ бал)-48->[въпрос 50]	0.338	0.034	9.823	0.000
(общ бал)-49->[въпрос 51]	0.073	0.038	1.903	0.057
(общ бал)-50->[въпрос 52]	0.246	0.036	6.757	0.000
(общ бал)-51->[въпрос 53]	0.301	0.035	8.533	0.000
(общ бал)-52->[въпрос 54]	0.032	0.039	0.821	0.412
(общ бал)-53->[въпрос 55]	0.149	0.038	3.949	0.000
(общ бал)-54->[въпрос 56]	-0.042	0.039	-1.075	0.282
(общ бал)-55->[въпрос 57]	0.358	0.034	10.581	0.000
(общ бал)-56->[въпрос 58]	0.322	0.035	9.250	0.000
(общ бал)-57->[въпрос 59]	0.417	0.032	12.951	0.000
(общ бал)-58->[въпрос 60]	0.332	0.035	9.597	0.000
(общ бал)-59->[въпрос 61]	0.427	0.032	13.421	0.000
(общ бал)-60->[въпрос 62]	0.202	0.037	5.426	0.000
(общ бал)-61->[въпрос 63]	0.328	0.035	9.452	0.000
(общ бал)-62->[въпрос 64]	0.014	0.039	0.360	0.719
(общ бал)-63->[въпрос 65]	0.072	0.038	1.861	0.063
(общ бал)-64->[въпрос 66]	0.208	0.037	5.614	0.000
(общ бал)-65->[въпрос 67]	0.067	0.039	1.729	0.084
(общ бал)-66->[въпрос 68]	0.148	0.038	3.922	0.000
(общ бал)-67->[въпрос 69]	0.163	0.038	4.315	0.000
(общ бал)-68->[въпрос 70]	0.153	0.038	4.040	0.000
(общ бал)-69->[въпрос 71]	0.276	0.036	7.712	0.000
(общ бал)-70->[въпрос 72]	0.129	0.038	3.382	0.001
(общ бал)-71->[въпрос 73]	0.119	0.038	3.132	0.002
(общ бал)-72->[въпрос 74]	0.024	0.039	0.634	0.526
(общ бал)-73->[въпрос 75]	0.087	0.038	2.253	0.024
(общ бал)-74->[въпрос 76]	0.213	0.037	5.758	0.000
(общ бал)-75->[въпрос 77]	0.465	0.031	15.212	0.000
(общ бал)-76->[въпрос 78]	0.239	0.037	6.541	0.000
(общ бал)-77->[въпрос 79]	0.293	0.035	8.250	0.000
(общ бал)-78->[въпрос 80]	0.292	0.036	8.213	0.000
(общ бал)-79->[въпрос 82]	0.563	0.027	20.947	0.000
(общ бал)-80->[въпрос 83]	0.391	0.033	11.857	0.000
(общ бал)-81->[въпрос 84]	0.559	0.027	20.724	0.000
(общ бал)-82->[въпрос 85]	0.403	0.033	12.371	0.000
(общ бал)-83->[въпрос 86]	0.447	0.031	14.326	0.000
(общ бал)-84->[въпрос 87]	0.477	0.030	15.788	0.000
(общ бал)-85->[въпрос 88]	0.410	0.032	12.676	0.000
(общ бал)-86->[въпрос 89]	0.471	0.030	15.478	0.000
(общ бал)-87->[въпрос 90]	0.508	0.029	17.474	0.000

Връзка	Оценки на модела			
	оценка на параметъра	стандартна грешка	Т-статистика	ниво на зна-чимост
(общ бал)-88->[въпрос 91]	0.492	0.030	16.594	0.000
(общ бал)-89->[въпрос 92]	0.408	0.032	12.576	0.000
(общ бал)-90->[въпрос 93]	0.331	0.035	9.560	0.000
(общ бал)-91->[въпрос 94]	0.386	0.033	11.646	0.000
(общ бал)-92->[въпрос 95]	0.326	0.035	9.377	0.000
(общ бал)-93->[въпрос 96]	0.340	0.034	9.912	0.000
(общ бал)-94->[въпрос 97]	0.596	0.025	23.460	0.000
(общ бал)-95->[въпрос 98]	0.341	0.034	9.941	0.000
(общ бал)-96->[въпрос 99]	0.459	0.031	14.905	0.000
(общ бал)-97->[въпрос 100]	0.221	0.037	5.998	0.000
(DELTA1)-->[въпрос 1]				
(DELTA2)-->[въпрос 2]				
(DELTA3)-->[въпрос 3]				
(DELTA4)-->[въпрос 4]				
(DELTA5)-->[въпрос 5]				
(DELTA6)-->[въпрос 6]				
(DELTA7)-->[въпрос 7]				
(DELTA8)-->[въпрос 8]				
(DELTA9)-->[въпрос 9]				
(DELTA10)-->[въпрос 10]				
(DELTA11)-->[въпрос 11]				
(DELTA12)-->[въпрос 12]				
(DELTA13)-->[въпрос 13]				
(DELTA14)-->[въпрос 14]				
(DELTA15)-->[въпрос 15]				
(DELTA16)-->[въпрос 16]				
(DELTA17)-->[въпрос 17]				
(DELTA18)-->[въпрос 18]				
(DELTA19)-->[въпрос 19]				
(DELTA20)-->[въпрос 20]				
(DELTA21)-->[въпрос 21]				
(DELTA22)-->[въпрос 22]				
(DELTA23)-->[въпрос 23]				
(DELTA24)-->[въпрос 24]				
(DELTA25)-->[въпрос 25]				
(DELTA26)-->[въпрос 26]				
(DELTA27)-->[въпрос 27]				
(DELTA28)-->[въпрос 28]				
(DELTA29)-->[въпрос 29]				
(DELTA30)-->[въпрос 30]				
(DELTA31)-->[въпрос 31]				
(DELTA32)-->[въпрос 32]				
(DELTA33)-->[въпрос 34]				
(DELTA34)-->[въпрос 35]				
(DELTA35)-->[въпрос 36]				
(DELTA36)-->[въпрос 37]				

Връзка	Оценки на модела			
	оценка на параметъра	стандартна грешка	Т-статистика	ниво на значимост
(DELTA37)-->[въпрос 38]				
(DELTA38)-->[въпрос 39]				
(DELTA39)-->[въпрос 40]				
(DELTA40)-->[въпрос 41]				
(DELTA41)-->[въпрос 43]				
(DELTA42)-->[въпрос 44]				
(DELTA43)-->[въпрос 45]				
(DELTA44)-->[въпрос 46]				
(DELTA45)-->[въпрос 47]				
(DELTA46)-->[въпрос 48]				
(DELTA47)-->[въпрос 49]				
(DELTA48)-->[въпрос 50]				
(DELTA49)-->[въпрос 51]				
(DELTA50)-->[въпрос 52]				
(DELTA51)-->[въпрос 53]				
(DELTA52)-->[въпрос 54]				
(DELTA53)-->[въпрос 55]				
(DELTA54)-->[въпрос 56]				
(DELTA55)-->[въпрос 57]				
(DELTA56)-->[въпрос 58]				
(DELTA57)-->[въпрос 59]				
(DELTA58)-->[въпрос 60]				
(DELTA59)-->[въпрос 61]				
(DELTA60)-->[въпрос 62]				
(DELTA61)-->[въпрос 63]				
(DELTA62)-->[въпрос 64]				
(DELTA63)-->[въпрос 65]				
(DELTA64)-->[въпрос 66]				
(DELTA65)-->[въпрос 67]				
(DELTA66)-->[въпрос 68]				
(DELTA67)-->[въпрос 69]				
(DELTA68)-->[въпрос 70]				
(DELTA69)-->[въпрос 71]				
(DELTA70)-->[въпрос 72]				
(DELTA71)-->[въпрос 73]				
(DELTA72)-->[въпрос 74]				
(DELTA73)-->[въпрос 75]				
(DELTA74)-->[въпрос 76]				
(DELTA75)-->[въпрос 77]				
(DELTA76)-->[въпрос 78]				
(DELTA77)-->[въпрос 79]				
(DELTA78)-->[въпрос 80]				
(DELTA79)-->[въпрос 82]				
(DELTA80)-->[въпрос 83]				
(DELTA81)-->[въпрос 84]				
(DELTA82)-->[въпрос 85]				

Връзка	Оценки на модела			
	оценка на параметъра	стандартна грешка	Т-статистика	ниво на зна-чимост
(DELTA83)-->[въпрос 86]				
(DELTA84)-->[въпрос 87]				
(DELTA85)-->[въпрос 88]				
(DELTA86)-->[въпрос 89]				
(DELTA87)-->[въпрос 90]				
(DELTA88)-->[въпрос 91]				
(DELTA89)-->[въпрос 92]				
(DELTA90)-->[въпрос 93]				
(DELTA91)-->[въпрос 94]				
(DELTA92)-->[въпрос 95]				
(DELTA93)-->[въпрос 96]				
(DELTA94)-->[въпрос 97]				
(DELTA95)-->[въпрос 98]				
(DELTA96)-->[въпрос 99]				
(DELTA97)-->[въпрос 100]				
(DELTA1)-98-(DELTA1)	0.821	0.027	30.301	0.000
(DELTA2)-99-(DELTA2)	0.821	0.027	30.332	0.000
(DELTA3)-100-(DELTA3)	0.593	0.030	19.773	0.000
(DELTA4)-101-(DELTA4)	0.813	0.027	29.666	0.000
(DELTA5)-102-(DELTA5)	0.894	0.023	39.491	0.000
(DELTA6)-103-(DELTA6)	0.718	0.030	24.006	0.000
(DELTA7)-104-(DELTA7)	0.943	0.018	53.848	0.000
(DELTA8)-105-(DELTA8)	0.765	0.029	26.391	0.000
(DELTA9)-106-(DELTA9)	0.909	0.021	42.628	0.000
(DELTA10)-107-(DELTA10)	0.928	0.019	47.959	0.000
(DELTA11)-108-(DELTA11)	0.961	0.015	65.140	0.000
(DELTA12)-109-(DELTA12)	0.933	0.019	49.695	0.000
(DELTA13)-110-(DELTA13)	0.862	0.025	34.619	0.000
(DELTA14)-111-(DELTA14)	0.976	0.012	83.486	0.000
(DELTA15)-112-(DELTA15)	0.873	0.024	36.155	0.000
(DELTA16)-113-(DELTA16)	0.970	0.013	74.119	0.000
(DELTA17)-114-(DELTA17)	0.873	0.024	36.093	0.000
(DELTA18)-115-(DELTA18)	0.862	0.025	34.617	0.000
(DELTA19)-116-(DELTA19)	0.861	0.025	34.441	0.000
(DELTA20)-117-(DELTA20)	0.899	0.022	40.494	0.000
(DELTA21)-118-(DELTA21)	0.808	0.028	29.280	0.000
(DELTA22)-119-(DELTA22)	0.797	0.028	28.408	0.000
(DELTA23)-120-(DELTA23)	0.957	0.015	62.359	0.000
(DELTA24)-121-(DELTA24)	0.957	0.015	62.317	0.000
(DELTA25)-122-(DELTA25)	0.817	0.027	29.954	0.000
(DELTA26)-123-(DELTA26)	0.754	0.029	25.795	0.000
(DELTA27)-124-(DELTA27)	0.941	0.018	52.963	0.000
(DELTA28)-125-(DELTA28)	0.982	0.010	96.609	0.000
(DELTA29)-126-(DELTA29)	0.636	0.030	21.014	0.000
(DELTA30)-127-(DELTA30)	0.879	0.024	36.955	0.000
(DELTA31)-128-(DELTA31)	0.928	0.019	47.896	0.000

Връзка	Оценки на модела			
	оценка на параметъра	стандартна грешка	Т-статистика	ниво на значимост
(DELTA32)-129-(DELTA32)	0.873	0.024	36.084	0.000
(DELTA33)-130-(DELTA33)	0.912	0.021	43.543	0.000
(DELTA34)-131-(DELTA34)	0.908	0.021	42.532	0.000
(DELTA35)-132-(DELTA35)	0.989	0.008	120.865	0.000
(DELTA36)-133-(DELTA36)	0.980	0.011	91.887	0.000
(DELTA37)-134-(DELTA37)	0.922	0.020	46.243	0.000
(DELTA38)-135-(DELTA38)	0.825	0.027	30.702	0.000
(DELTA39)-136-(DELTA39)	0.767	0.029	26.481	0.000
(DELTA40)-137-(DELTA40)	0.558	0.030	18.895	0.000
(DELTA41)-138-(DELTA41)	0.871	0.024	35.768	0.000
(DELTA42)-139-(DELTA42)	0.818	0.027	30.093	0.000
(DELTA43)-140-(DELTA43)	0.721	0.030	24.170	0.000
(DELTA44)-141-(DELTA44)	0.815	0.027	29.798	0.000
(DELTA45)-142-(DELTA45)	0.999	0.002	509.857	0.000
(DELTA46)-143-(DELTA46)	0.865	0.025	34.943	0.000
(DELTA47)-144-(DELTA47)	0.908	0.021	42.555	0.000
(DELTA48)-145-(DELTA48)	0.886	0.023	38.078	0.000
(DELTA49)-146-(DELTA49)	0.995	0.006	176.489	0.000
(DELTA50)-147-(DELTA50)	0.939	0.018	52.423	0.000
(DELTA51)-148-(DELTA51)	0.909	0.021	42.786	0.000
(DELTA52)-149-(DELTA52)	0.999	0.002	407.434	0.000
(DELTA53)-150-(DELTA53)	0.978	0.011	86.439	0.000
(DELTA54)-151-(DELTA54)	0.998	0.003	311.458	0.000
(DELTA55)-152-(DELTA55)	0.872	0.024	35.885	0.000
(DELTA56)-153-(DELTA56)	0.896	0.022	39.995	0.000
(DELTA57)-154-(DELTA57)	0.826	0.027	30.790	0.000
(DELTA58)-155-(DELTA58)	0.890	0.023	38.804	0.000
(DELTA59)-156-(DELTA59)	0.817	0.027	30.010	0.000
(DELTA60)-157-(DELTA60)	0.959	0.015	64.020	0.000
(DELTA61)-158-(DELTA61)	0.893	0.023	39.290	0.000
(DELTA62)-159-(DELTA62)	1.000	0.001	928.110	0.000
(DELTA63)-160-(DELTA63)	0.995	0.006	180.415	0.000
(DELTA64)-161-(DELTA64)	0.957	0.015	62.042	0.000
(DELTA65)-162-(DELTA65)	0.996	0.005	194.132	0.000
(DELTA66)-163-(DELTA66)	0.978	0.011	87.003	0.000
(DELTA67)-164-(DELTA67)	0.974	0.012	79.412	0.000
(DELTA68)-165-(DELTA68)	0.977	0.012	84.575	0.000
(DELTA69)-166-(DELTA69)	0.924	0.020	46.659	0.000
(DELTA70)-167-(DELTA70)	0.983	0.010	100.376	0.000
(DELTA71)-168-(DELTA71)	0.986	0.009	108.136	0.000
(DELTA72)-169-(DELTA72)	0.999	0.002	527.836	0.000
(DELTA73)-170-(DELTA73)	0.993	0.007	149.385	0.000
(DELTA74)-171-(DELTA74)	0.955	0.016	60.611	0.000
(DELTA75)-172-(DELTA75)	0.783	0.028	27.518	0.000
(DELTA76)-173-(DELTA76)	0.943	0.017	53.975	0.000
(DELTA77)-174-(DELTA77)	0.914	0.021	44.028	0.000

Връзка	Оценки на модела			
	оценка на параметъра	стандартна грешка	Т-статистика	ниво на зна-чимост
(DELTA78)-175-(DELTA78)	0.915	0.021	44.198	0.000
(DELTA79)-176-(DELTA79)	0.683	0.030	22.610	0.000
(DELTA80)-177-(DELTA80)	0.847	0.026	32.867	0.000
(DELTA81)-178-(DELTA81)	0.687	0.030	22.746	0.000
(DELTA82)-179-(DELTA82)	0.837	0.026	31.841	0.000
(DELTA83)-180-(DELTA83)	0.800	0.028	28.666	0.000
(DELTA84)-181-(DELTA84)	0.773	0.029	26.848	0.000
(DELTA85)-182-(DELTA85)	0.831	0.027	31.275	0.000
(DELTA86)-183-(DELTA86)	0.779	0.029	27.202	0.000
(DELTA87)-184-(DELTA87)	0.742	0.030	25.161	0.000
(DELTA88)-185-(DELTA88)	0.758	0.029	25.995	0.000
(DELTA89)-186-(DELTA89)	0.833	0.026	31.457	0.000
(DELTA90)-187-(DELTA90)	0.891	0.023	38.930	0.000
(DELTA91)-188-(DELTA91)	0.851	0.026	33.316	0.000
(DELTA92)-189-(DELTA92)	0.894	0.023	39.548	0.000
(DELTA93)-190-(DELTA93)	0.884	0.023	37.803	0.000
(DELTA94)-191-(DELTA94)	0.645	0.030	21.270	0.000
(DELTA95)-192-(DELTA95)	0.884	0.023	37.714	0.000
(DELTA96)-193-(DELTA96)	0.789	0.028	27.900	0.000
(DELTA97)-194-(DELTA97)	0.951	0.016	58.379	0.000